HIGH RESOLUTION RELATIVE DETECTION VIA INFERENCE OF IDENTICAL BY DESCENT SHARING OF SAMPLE ANCESTORS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Monica Denise Ramstetter

May 2017

HIGH RESOLUTION RELATIVE DETECTION VIA INFERENCE OF IDENTICAL BY DESCENT SHARING OF SAMPLE ANCESTORS

Monica Denise Ramstetter, Ph.D.

Cornell University 2017

Inferring relatedness from genomic data is an essential component of genetic association studies, population genetics, forensics, and genealogy. Due to the random nature of Mendelian inheritance, variance in the amount of the genome shared identically between two individuals of a certain degree of relatedness can be high, making relatedness inference difficult. While numerous methods exist for performing such inference, thorough evaluation of these methods in real data has been lacking. We assessed 11 state-of-the-art relatedness inference methods using a dataset with 2,485 individuals contained in several large pedigrees that span up to six generations. Overall, the methods have high accuracy (93%-99%) when reporting first and second degree relationships, but less than 60% accuracy for fifth degree relationships. We considered a composite method built off the three methods with highest accuracy in our analysis (ERSA 2.0, IBDseq, and Refined IBD) and applied it to the SAMAFS, HapMap3, and Weill Cornell Qatari datasets, finding numerous unreported relationships in all three datasets. Building on the insights from our analysis of methods, we developed DRUID—Deep Relatedness Utilizing Identity by Descent—a method that works by inferring the identical by descent (IBD) sharing profile of an ungenotyped ancestor of a set of close relatives. DRUID combines relatedness signals among

multiple samples to effectively remove one or more generations of distance between a set of relatives, leading to substantial accuracy improvements compared to other methods.

BIOGRAPHICAL SKETCH

Monica Ramstetter graduated in June 2011 from the University of California at Irvine (UCI) with a B.S. in mathematics with honors, specializing in applied and computational math, and a B.A. in quantitative economics. At UCI, she assisted with research in two biology-based labs: that of Dr. Sabee Molloi and that of Dr. Adrianna Briscoe. In Dr. Briscoe's lab, she received numerous grants for her research and published her first co-authored paper on butterfly wing coloration and visual systems. She attended graduate school at Cornell University from 2011 through 2017 under the guidance of both Dr. Jason Mezey and Dr. Amy Williams, receiving the Presidential Life Sciences Fellowship award and honorable mention for the National Science Foundation (NSF) Graduate Research Fellowship Program (GRFP) her first year, followed by a four-year GRFP fellowship the following year. She received an NSF scholarship to attend a three week workshop, Rice: Research to Production, at the International Rice Research Institute in the Philippines. She also completed a summer internship with Monsanto in St. Louis, MO in 2016. Although she worked on numerous projects of varying topics throughout her graduate career, her dissertation research focuses on inferring genetic relatedness between individuals.

This document is dedicated to the family, friends, faculty, and hundreds of pounds of coffee beans who supported my efforts.

You have my most sincere gratitude.

Science is the best idea humans have ever had.

The more people who embrace that idea, the better.

–Bill Nye

ACKNOWLEDGEMENTS

Family

Mom, thank you for providing encouragement throughout my graduate career. You are such an inspirational person and I strive to be as strong as you each day.

Grandma, thank you for all never-ending generosity. Being able to visit you at often stressful times provided much needed joy and laughter in my life. I'm fairly certain my mom and I inherited our curious natures from you, and for that, I am so very, very grateful.

Harry, thank you for working with a long-distance relationship. I know it was hard, but I am so exceptionally happy to be reunited with you.

Harry's family, thank you for remembering me with cards around the holidays while I was on the opposite coast. It was so comforting to know I was never forgotten.

Rick and Erica, thank you for being the goofy people you are. Your visits never failed to bring me smiles and laughter. Roan, I cannot wait to meet you!

Academia

Jason, you had faith in my abilities when no one else, including myself, did. I cannot express enough gratitude for that.

Amy, your cheerfulness and excitement about science brought back the love of science I feared I was losing, and your support and guidance have been so unbelievably helpful—I cannot thank you enough for all of that.

Andy and Jacob, thank you for your understanding and your coping with my often-changing situations, and thank you for your guidance in leading me toward earning a PhD. You have been wonderful to have on my committee.

Susan, thank you for the numerous opportunities you provided me, particularly the Rice to Research workshop. Traveling outside of North America and meeting so many interesting, intelligent researchers around the globe was truly inspirational and life-changing.

Sue, your never-fading smile and your eagerness to try to help solve whatever problem came up was always so very appreciated.

Labmates and other friends made at Cornell, each of you made such a difference in my life. You are all incredibly smart yet humble people, and I loved and miss all of our day-to-day banter.

Other

To the National Science Foundation and Cornell University: the various funding and opportunities you awarded and provided me have changed my life in so many ways. From giving me with the ease of mind in knowing that I never needed to worry about

whether my PI could continue supporting me, to giving me the opportunity to travel around the world, your generosity has sculpted me into a researcher with so many interesting experiences who is eager to help change the world for the better.

The material here is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1144153.

.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INFERRING GENETIC RELATEDNESS BETWEEN INDIVIDUALS

In 1990, the scientific community was excited to launch the Human Genome Project, a 13 year endeavor to determine the DNA sequence of a human genome[3]. Shortly after its completion, many projects such as the International HapMap Project[4] and the 1000 Genomes Project[5] began, quickly revealing how improving technology could allow for the sequencing of thousands of individuals. Today, datasets easily contain just as many, if not more, individuals, with the UK leading an endeavor to sequence the genomes of 100,000 individuals (the 100,000 Genomes Project). The ever-decreasing cost — both with respect to money and time — of sequencing means that we can someday expect to see datasets with sample sizes in the millions.

Larger sample sizes open doors for new discoveries. In order to understand, characterize, and better manage or even cure diseases, disease-causing genetic variants must be discovered. However, these discoveries heavily depend on increasing the amount and quality of genetic data available for analyses. The "Common Disease-Rare Variant Hypothesis" speculates that some common diseases may be polygenic, suggesting the genetic cause of that disease varies between individuals in the population, with each of these genetic causes being rare within the population[6]. In order to discover such rare variants with low penetrance as suggested by this hypothesis, large numbers of samples must be analyzed. Though the strong drive to increase sample sizes will lead to better characterization of disease, it comes with a price:

the amount of data to be analyzed can easily become unwieldy[7], and certain pieces of information critical for proper interpretation of results such as close relatedness between pairs of individuals can easily go misreported or unreported.

1.1 The Importance of Detecting Relatedness between Samples

The inference of relatedness has a wide application of topics: in genetic association studies[8–10], it is crucial to account for close relatives in order to avoid biased genetic signals and spurious associations; in linkage analysis[11–13], relationships must be properly specified; in forensic genetics[14–16], it is used as a tool to assist in determining relatives of missing persons, victims of disasters, and criminals (such as with 'familial searching'[17]); in population genetic analyses[18–20], it is needed to account for or remove relatives to avoid bias. In genome-wide association studies in particular, population structure (substantial differences in ancestry in a sample set, possibly from groups of individuals sharing more recent ancestors than expected in a random-mating population, that are reflected in the genomes of the individuals) and cryptic relatedness (unreported close relatedness between sets of sampled individuals) alter the genome-wide distribution per-SNP p-values, inflating false-positive associations if not accounted for[21–24]. The most common method used to account for population structure is the inclusion of principal component (PC) information from

principal components analysis (PCA) into the model of interest. This method considers the top principal components to be continuous axes of variation which reflect ancestry-based genetic variation in the dataset and corrects for those[25]. However, if there exists a smaller number of otherwise unaccounted for related individuals in the data, the top PCs may reflect familial relatedness rather than population structure[26]. Thus, the proper analysis of GWAS data hinges on researchers' abilities to detect and account for possible close relatives.

Even outside the scientific community, relatedness inference has grown popular: companies such as 23andMe and AncestryDNA advertise their ability to find and report relatives, allowing individuals from the general public to explore their ancestry and genealogy. Further, marriage and inheritance laws are sometimes based on the degree of relationship among family members. Both within and outside the realm of scientific research, relatedness inference proves to be a necessity, especially in an age where personalized medicine based on genetic studies is becoming more feasible[27].

As more individuals are sampled, especially when they are sampled from the same population, the probability of cryptic relatedness grows: it has been suggested that random pairs of individuals in Europe are fairly likely to share a common ancestor within the past 1,000 years[28]. Even in long-term studies that included extensive quality control measures such as HapMap, unreported close relationships have been found[29], making it necessary for researchers to detect and account for possibly unreported close relatives in their data. Further, errors in reported relationships may

exist[1,29,30], reinforcing researchers' need to check for close relatives, regardless of whether reported relationships are available or not.

1.2 Relatedness Inference

In 1866, Mendel described what is now called Mendel's first law: the descent of DNA from one individual to his/her offspring[31]. This fundamental law of inheritance suggests that in a diploid individual, at each location in the genome, one homologous copy of DNA is passed at random to the offspring gamete. This random passing of DNA is the basis for relatedness inference as it provides a framework for understanding how closely related two individuals are based on the DNA each one inherited.

1.2.1 DNA Inheritance, Recombination, and Identity by Descent

Large segments of DNA are passed between generations, and these passed segments, though broken up by recombination over time, can give insight to the proportion of genome which is shared identical by descent (IBD) between two individuals, es-

sentially giving insight into their shared ancestry[32]. IBD describes the case when two or more individuals inherit a segment of DNA from the same recent common ancestor. For example, we expect a parent-child pair to share 1/2 of their genomes IBD since that parent, the recent common ancestor in this case, passes down 50% of his/her genes to the child. Mathematically, IBD is the probability that we observe a haplotype in one individual that is not independent of observing a haplotype in another individual. However, any two individuals in a finite population are related in that they necessarily share a common founding ancestor in the past. Therefore, probabilities of IBD sharing are defined relative to a reference point some number of generations in the past at which all ancestors are assumed to be unrelated.

IBD is the backbone to relatedness inference as it evidences similar ancestries between individuals via their inherited DNA. Though we can describe expected amounts of IBD sharing between pairs of individuals of certain relationship types, the probabilistic nature of Mendelian inheritance and recombination during meiosis creates large relative variance around these amounts[15,33]. This variance and the exponential decrease of proportion of genome shared IBD with each increase in degree of relatedness make relatedness inference difficult. Though on average, more closely related individuals are IBD across a larger portion of their genomes, the proportion of genome shared IBD and the actual pedigree relationship can vary: as the number of reproductive events which separate two individual increases, so does the number of random transmissions from parents to children, creating greater variation in the proportion of the genome that is inherited from the common ancestor.

1.2.2 Identity by Descent

When inbreeding is ignored, there exists three classifications of IBD. If a pair of individuals shares both haplotype copies at a locus, and those each of those haplotypes was inherited from the same common ancestor (with respect to a reference population some number of generations in the past), we say that they share that locus IBD2; if they share only one haplotype copy which was inherited from the same common ancestor at the locus, we say they share that locus IBD1; if they share no haplotype copy that was inherited from the same common ancestor at the locus, we say they share that locus IBD0. For example, a parent is expected to be IBD1 with his/her child at all regions of the genome as exactly one of the parent's haplotypes is passed to the child throughout the region. Differentiating between IBD0, IBD1, and IBD2 allows researchers to distinguish between different relationship types within a given degree of relatedness, such as parent-child (E(proportion of genome shared IBD0) = 0, E(proportion of genome shared IBD1) = 1, E(proportion of genome shared IBD2) = 0) and siblings (E(proportion of genome shared IBD0) = $\frac{1}{4}$, E(proportion of genome shared IBD1) = $\frac{1}{2}$, E(proportion of genome shared IBD2) = $\frac{1}{4}$).

In general, the inference of relatedness between individuals does not require distinguishing between maternal and paternal haplotypes — however, when attempting to account for inbreeding, there are nine possible states of IBD sharing as shown in Figure 1.1. For the standard case of no inbreeding, which is the focus of this

thesis, there remain only three IBD-sharing configurations: Δ_7, Δ_8, and Δ_9, called Jacquard coefficients[34], where $\sum_{i=7}^{9} \Delta_i = 1$. Here, Δ_7 represents the probability that at a random locus, the two individuals are IBD0, Δ_8, the probability that the two individuals are IBD1 at a random locus, and Δ_9, the probability that the two individuals are IBD2 at a random locus.

The task of inferring IBD regions and/or proportions of the genome shared IBD is not trivial and requires either leveraging information from independent SNPs via their allele frequencies or inferring haplotypes and using statistical or heuristic approaches to determine whether two identical haplotype segments are IBD. Since IBD tracts are broken up by recombination during meiosis, shorter IBD tracts are likely to arise from a more distant ancestor and longer IBD tracts are likely to descend from a recent ancestor. In practice, distinguishing IBD segments from chance sharing of haplotypes involves analyzing the population frequency of the haplotype. A shared haplotype with low population frequency provides evidence of more recent relatedness due to its rarity. Thus, both the lengths of shared haplotypes and the frequency of shared haplotypes in the population can be used to infer regions of IBD sharing.

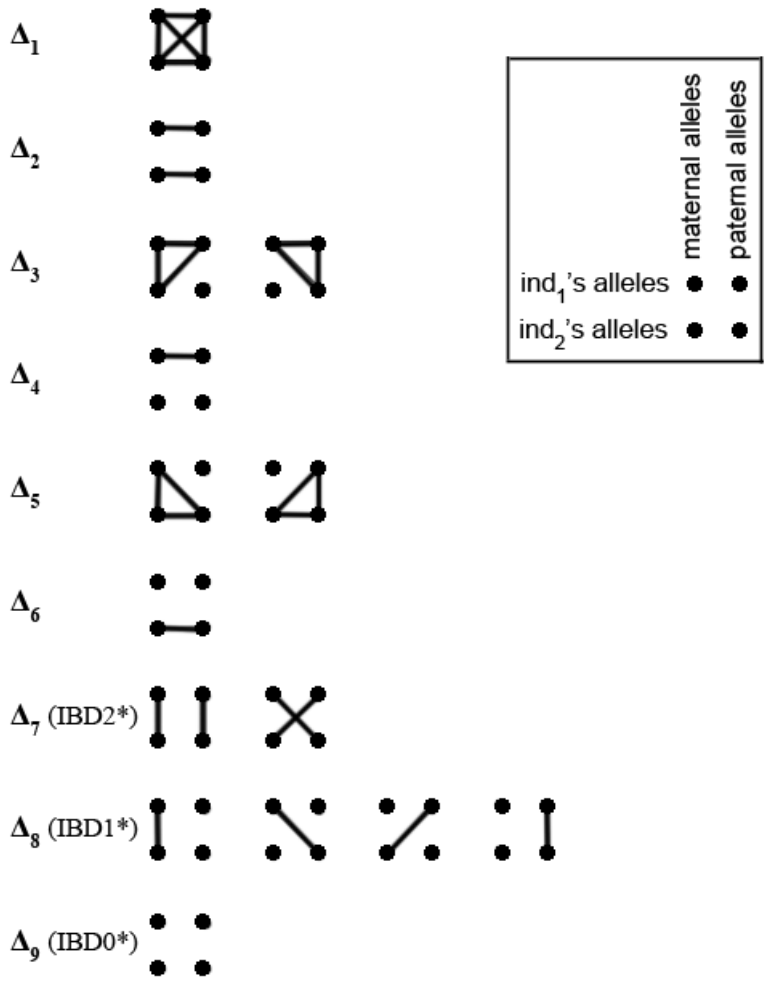

Figure 1.1: Jacquard coefficients and their relation to haplotype sharing between two individuals. *In cases when no inbreeding is present, only $\Delta_7, \Delta_8, and \Delta_9$ are possible, and these represent the probabilities of a locus being shared IBD2, IBD1, and IBD0, respectively.

1.2.3 The Kinship Coefficient and Degree of Relatedness

Determining the probability of two individuals' genotypes at a locus when their relationship is known is straightforward[13,35–37], but the reverse—determining the proba-

bility of a relationship when their genotypes are known—is much more difficult. For example, given an individual s who is homozygous for an allele which has frequency 0.1 in the population, the probability that someone unrelated to s is also homozygous for the same allele is simply $0.1 \times 0.1 = 0.01$, but the probability a sibling of s is also homozygous for the same allele given s is homozygous is increased. This is due to the additional information about the genotypes of the parents of s and hence about that of the sibling through s's genotype: because s is homozygous, the parents must each be either heterozygous or homozygous for the allele, increasing the probability that s's sibling is homozygous for the allele. If we were not aware these two individuals were siblings and we found they were both homozygous for the same allele, we would be unable to say based off that single genotype whether the two individuals are siblings since unrelated individuals can also be homozygous for the same allele. By instead considering loci across the genome that are inferred to be IBD between two individuals, we can attempt to estimate the level of relatedness between these individuals. The kinship coefficient, ϕ, is a function of the genome-wide proportion of IBD-sharing between a pair of individuals, i and j, that denotes the probability that two randomly selected alleles at a locus are IBD for individuals i and j. It can be conveniently calculated as $\phi_{ij} = \frac{p_{ij}^{(1)}}{4} + \frac{p_{ij}^{(2)}}{2}$, where $p_{ij}^{(1)}$ and $p_{ij}^{(2)}$ denote the proportion of their genomes that individuals i, j share IBD1 and IBD2, respectively. These $p_{ij}^{(1)}$ and $p_{ij}^{(2)}$ are simply the sum of the genetic lengths of the IBD1 and IBD2 segments, respectively, between samples i, j divided by the total genetic length of the genome analyzed. (Note if $i = j$, then $\phi_{ii} = \frac{1}{2}(1 + i)$ where f_i is the kinship coefficient between the parents of i which is equivalent to the inbreeding coefficient

of individual i.)

The kinship coefficient is the same as the coefficient of coancestry defined by Sewall Wright in 1922[38]. The estimated kinship coefficient can be used to determine an inferred degree of relatedness, or a measure of relatedness between two individuals, based previously reported simulation-based ranges[1] (see Table 1.1): for example, first degree relatives include parent-child pairs and full-sibling pairs, and second degree relatives include grandparent-grandchild pairs, avuncular (aunt/uncle and niece/nephew) pairs, double-cousins, and half-siblings. Relatedness degrees are based on expected proportion of genome shared IBD for a pair of individuals, hence why parent-child pairs and sibling pairs are both considered first degree relationships (expected proportion of genome shared IBD for these types equals 50%), why avuncular pairs, grandparent-grandchild pairs, double-cousins, and half-siblings are considered second degree relationships (expected proportion of genome shared IBD for these types equals 25%), and so on. However, degrees of relatedness are only defined for expected proportions of genomes shared IBD which take the values in $\frac{1}{2^x}$ for $x \in \{1, 2, 3...\}$, making some cases of relatedness not directly map to a degree of relatedness. For example, three-quarter-siblings, or individuals who share one parent in common and whose unshared parents have a mean coefficient of relatedness of 50% (consistent with these parents being full-siblings), have a higher expected kinship coefficient than half-siblings but a lower expected kinship coefficient than full-siblings, meaning their level of relatedness falls between the first and second degree classifications.

			Expected				Accepted range for:	
Relationship	Degree	# Meiosis	IBD0	IBD1	IBD2	ϕ	ϕ	P(IBD=0)
Parent-child	1	1	0	1	0	$\frac{1}{2^2}$	$(\frac{1}{2^{3/2}}, \frac{1}{2^{1/2}}]$	< 0.1
Full siblings (not MZ twin)	1	2 (×2)	1/4	1/2	1/4	$\frac{1}{2^2}$	$(\frac{1}{2^{3/2}}, \frac{1}{2^{1/2}}]$	$[0.1, 0.365)$
Grandparent	2	2	1/2	1/2	0	$\frac{1}{2^3}$	$(\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}}]$	$[0.365, 1-\frac{1}{2^{3/2}})$
Avuncular	2	3 (×2)	1/2	1/2	0	$\frac{1}{2^3}$	$(\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}}]$	$[0.365, 1-\frac{1}{2^{3/2}})$
Double-cousins	2	4 (×4)	9/16	3/8	1/16	$\frac{1}{2^3}$	$(\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}}]$	$[0.365, 1-\frac{1}{2^{3/2}})$
Half-sibling	2	2	1/2	1/2	0	$\frac{1}{2^3}$	$(\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}}]$	$[0.365, 1-\frac{1}{2^{3/2}})$
First Cousin	3	4 (×2)	3/4	1/4	0	$\frac{1}{2^4}$	$(\frac{1}{2^{7/2}}, \frac{1}{2^{5/2}}]$	$[1-\frac{1}{2^{3/2}}, 1-\frac{1}{2^{5/2}})$
Double half-cousins	3	5 (×2)	23/32	7/32	1/64	$\frac{1}{2^4}$	$(\frac{1}{2^{7/2}}, \frac{1}{2^{5/2}}]$	$[1-\frac{1}{2^{3/2}}, 1-\frac{1}{2^{5/2}})$
Great-grandparent	3	3	3/4	1/4	0	$\frac{1}{2^4}$	$(\frac{1}{2^{7/2}}, \frac{1}{2^{5/2}}]$	$[1-\frac{1}{2^{3/2}}, 1-\frac{1}{2^{5/2}})$
Grand-avuncular	3	4 (×2)	3/4	1/4	0	$\frac{1}{2^4}$	$(\frac{1}{2^{7/2}}, \frac{1}{2^{5/2}}]$	$[1-\frac{1}{2^{3/2}}, 1-\frac{1}{2^{5/2}})$
Half-avuncular	3	4	3/4	1/4	0	$\frac{1}{2^4}$	$(\frac{1}{2^{7/2}}, \frac{1}{2^{5/2}}]$	$[1-\frac{1}{2^{3/2}}, 1-\frac{1}{2^{5/2}})$
First cousin once removed	4	5 (×2)	7/8	1/8	0	$\frac{1}{2^5}$	$(\frac{1}{2^{9/2}}, \frac{1}{2^{7/2}}]$	$[1-\frac{1}{2^{5/2}}, 1-\frac{1}{2^{7/2}})$
Great-great-grandparent	4	4	7/8	1/8	0	$\frac{1}{2^5}$	$(\frac{1}{2^{9/2}}, \frac{1}{2^{7/2}}]$	$[1-\frac{1}{2^{5/2}}, 1-\frac{1}{2^{7/2}})$
Great-grand-avuncular	4	5 (×2)	7/8	1/8	0	$\frac{1}{2^5}$	$(\frac{1}{2^{9/2}}, \frac{1}{2^{7/2}}]$	$[1-\frac{1}{2^{5/2}}, 1-\frac{1}{2^{7/2}})$
Half-grand-avuncular	4	5	7/8	1/8	0	$\frac{1}{2^5}$	$(\frac{1}{2^{9/2}}, \frac{1}{2^{7/2}}]$	$[1-\frac{1}{2^{5/2}}, 1-\frac{1}{2^{7/2}})$
First cousin twice removed	5	6 (×2)	15/16	1/16	0	$\frac{1}{2^6}$	$(\frac{1}{2^{11/2}}, \frac{1}{2^{9/2}}]$	$[1-\frac{1}{2^{7/2}}, 1-\frac{1}{2^{9/2}})$
Second cousin	5	6 (×2)	15/16	1/16	0	$\frac{1}{2^6}$	$(\frac{1}{2^{11/2}}, \frac{1}{2^{9/2}}]$	$[1-\frac{1}{2^{7/2}}, 1-\frac{1}{2^{9/2}})$
GGG-grandparent	5	5	15/16	1/16	0	$\frac{1}{2^6}$	$(\frac{1}{2^{11/2}}, \frac{1}{2^{9/2}}]$	$[1-\frac{1}{2^{7/2}}, 1-\frac{1}{2^{9/2}})$

Table 1.1: For a range of relationship types, the corresponding degree of relatedness of the individuals; the number of meioses that separate them, with (×2) indicating samples that are related along two lines of descent (such as full-siblings) that have the listed meiotic distance on both lines; proportions of the genome that are expected to be IBD0, IBD1, and IBD2 between the samples; and expected kinship coefficient ϕ. For inferring a degree of relatedness from either a kinship coefficient or a proportion of genome shared IBD0, the range of values that map to the given degree are listed (these ranges taken from Manichaikul *et al.*[1]). The list does not include all possible relationship types for the degrees of relatedness listed.

1.2.4 Difficulties of Relatedness Inference

It has previously been shown via simulations of IBD-sharing amongst various relative types that the credible interval for realized IBD of third degree relatives (e.g., first cousins) overlaps that of fourth degree relatives (e.g., half-cousins)[39,40]; this trend of overlap becomes more extreme for higher degrees, as well. Although the standard deviation of the proportion of IBD shared decreases as two individuals become less related, the coefficient of variation increases[39,41], resulting in the overlapping distributions. This overlap complicates one's ability to discriminate between different degrees of relatedness.

Aside from overlapping distributions of IBD sharing for different degrees of relatedness, SNP-based measures of genome similarity depend on the minor allele frequencies (MAFs) of the SNP set, but these MAFs are generally estimated from the data or a reference panel and are therefore dependent on the choice of SNP genotyping technology and the quality control procedures applied to the data[40]. When MAFs are estimated from the data and used in the estimation of relatedness, the inferred level of relatedness can be biased downward[40].

A further issue arises for allele frequency-based methods when allele frequencies are estimated from the sample data as opposed to from the full population. Any two individuals in a finite population are related in that they must share a common

ancestor. Pairs of individuals from the same population likely share a more recent common ancestor than those from different populations, making them more closely related than pairs of individuals between two populations even though this recent common ancestor may be many generations back. Allele frequencies in different subpopulations may therefore vary, possibly resulting in individuals within the same subpopulation appearing more closely related than they truly are when more than one subpopulation is included in the sample set.

1.3 Current Methods for Relatedness Inference

Methods for relatedness inference may attempt to distinguish between alternative relationships or estimate the degree of relatedness between pairs of individuals. The differentiation between alternative relationships can be done via likelihood ratios[35]. The likelihood of a relationship R is

$$L(R) = \prod_{j=1}^{p} \Pr(G_{1,j}, G_{2,j}|R) \tag{1.1}$$

where p is the number of loci and $G_{i,j}$ is the genotype of individual i at locus j. We obtain the probability of genotypes given a relationship type by noting that

$$\Pr(G_{1,j}, G_{2,j}|R) = \sum_{z} \Pr(G_{1,j}, G_{2,j}|z) \times \pi(z|R) \tag{1.2}$$

where $\pi(z|R)$ is the probability of IBD state z at a locus, $z \in \{0, 1, 2\}$ (assuming no inbreeding).

Many methods[13,35–37] are motivated by such likelihood models. However, as the number of independently segregating loci in the human genome is limited, this approach cannot extend far beyond inferring parent-offspring, sibling, and half-sibling pairs[32].

Methods for detecting recent ancestral relatedness between pairs of individuals generally make use of genome-wide IBD estimates or similar estimates to infer pairwise relationships. Some of these methods are based on Hidden Markov Models (HMMs) and only consider a small number of relationship types[13,42–44], whereas others may utilize estimates of probability of genome shared IBD0, IBD1, or IBD2,[45,46] or the kinship coefficient (ϕ)[1,47,48], also known as Wright's coefficient of coancestry[38], which can be described as the probability that an allele selected randomly from one individual and an allele selected randomly from the same autosomal locus of another individual are IBD.

Allele frequency-based methods of relatedness inference offer the advantage of generally being more computationally efficient than methods that leverage or infer IBD segments, making them convenient to apply in large datasets. However, they suffer from challenges in allele frequency estimation as already noted (Section 1.2.4). They further utilize less data and may not perform as well as IBD-based methods (Section 2.1). IBD segment-based methods offer the advantage of not requiring allele frequencies in the model, but have their own challenge: shorter IBD regions are harder to infer, meaning distantly-related pairs of individuals will have little or no

IBD detected[40], possibly adversely impacting their accuracy.

In this thesis, we seek to understand state-of-the-art relatedness inference methods and to improve upon these methods, particularly in the context of large datasets that contain many close relatives. In Chapter 2, we provide a rigorous evaluation of 11 methods that can scale to large study sizes. In Chapter 3, we create a composite method from the top-performing methods in Chapter 2 and apply it to three datasets. Chapter 4 introduces our novel approach to the problem of inferring relatedness, DRUID, which combines IBD signals across sets of close relatives to better infer more distant relatedness. And finally, Chapter 5 provides concluding remarks.

CHAPTER 2

COMPARISON AND AGGREGATION OF METHODS FOR RELATEDNESS INFERENCE

Presented here is a rigorous evaluation of 11 state-of-the-art methods that can scale to large study sizes, including seven that directly infer genome-wide relatedness measures[1,46,48–52] and four IBD segment detection methods[53–56] that were utilized to infer these quantities. To assess each of these methods, we used SNP array genotypes from Mexican American individuals contained in large pedigrees from the San Antonio Mexican American Family Studies (SAMAFS)[57–59]. Our analysis sample included 2,485 individuals genotyped at 521,184 SNPs within pedigrees that span up to six generations with genotype data from as many as five generations of individuals. Given this large sample, including 13 pedigrees with >50 individuals (Figure 2.1), numerous close relatives exist, and we used these to evaluate each of the inference methods. In particular, there are >4,500 pairs of individuals within each of the first through fifth degree relatedness classes that we evaluated, and we further considered more than three million pairs of individuals that are in distinct pedigrees and hence assumed unrelated (Table 2.1). Prior analyses of relatedness inference methods considered either simulated data[1,48,50–52]—which may not fully capture the complexities of real data—or used small sample sizes[1,48,52,60]. Our analysis using real data for large numbers of up to fifth degree relatives provides a comprehensive evaluation of these relatedness inference methods.

We focused on SNP array data from the San Antonio Mexican American Family Studies[57–59] (SAMAFS). The original set of 2,490 samples were genotyped via one of the following Illumina arrays: the Human660W, Human1M, Human1M-Duo, or both the HumanHap500 and the HumanExon510S array which together provide roughly the same content as the Human1M array. We began by using data that had quality control measures carried out in a prior study[61]. In brief, sites with non-Mendelian errors were set to missing and BWA[62] v0.7.5a-r405 was used to map the SNP array probe sequences to human reference sequence GRCh37. Only SNPs with probe sequences that aligned with no mismatches at exactly one genomic position and that do not align to any other location with either zero or one mismatches were kept. We omitted SNPs for which any of the following was true: (1) multiple probes aligned to the same location (we retained only one SNP for any location), (2) dbSNP reported either more than two variants or had non-simple alleles (i.e., not A/C/G/T), (3) the raw genotype data had differing alleles as compared to those reported in the manifest files, (4) the manifest file listed SNP alleles that differed from those in dbSNP at the aligned location, (5) dbSNP listed the SNP as having multiple locations or as 'suspected', (6) the SNP location is outside the 'accessible genome' as reported by the 1000 Genomes Project[63], (7) the site occurs in a region that is segmentally duplicated with a Jukes-Cantor K-value of $<2\%$, or (8) the site occurs within a total of 17 Mb of the genome that received excess reads aligned using 1000 Genome Project data[64].

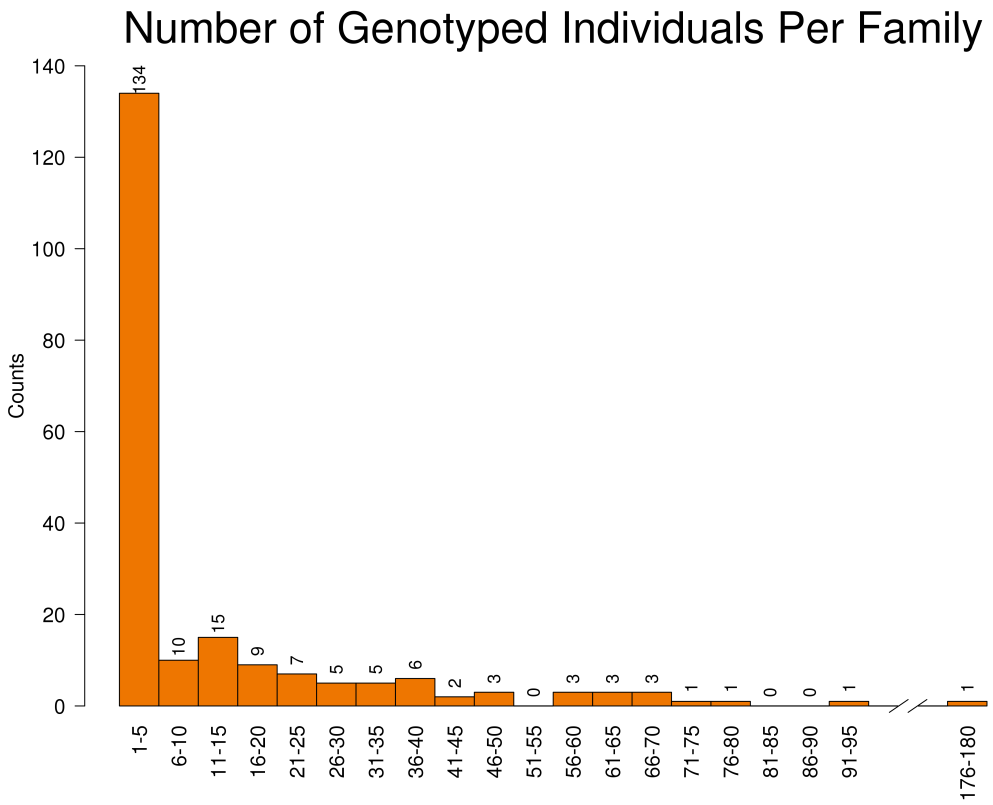

Figure 2.1: Histogram of the number of genotyped individuals within pedigrees in the SAMAFS dataset.

Following these procedures, we filtered the dataset to include SNPs with less than 2% missingness and individuals with less than 10% missingness. This resulted in data for a total of 2,485 individuals at 521,184 SNPs that overlap between the two types of arrays and are of high quality. To run methods that require independent SNPs, we used the PLINK command `--indep-pairwise 1000 25 0.25` which uses a sliding window method that considers blocks of 1,000 SNPs and removes SNPs with $r^2 > 0.25$, afterward shifting the window by 25 SNPs. This process yielded a

Degree	Number of Pairs
1	4,969
2	6,590
3	8,244
4	7,950
5	4,501
Unrelated	3,025,035
Total	3,057,289

Table 2.1: Numbers of pairs of individuals from the SAMAFS dataset reported to have relatedness between first and fifth degree and counts of unrelated pairs used for the evaluation. Only individuals from distinct pedigrees are considered unrelated.

total of 140,314 SNPs filtered to remove linkage disequilibrium (LD).

For the ADMIXTURE analyses described in Section 2.2.3, we merged the above LD-pruned SAMAFS dataset with HapMap phase 3 (HapMap3) samples[65] and again filtered to include SNPs with less than 2% missingness from the combined dataset. This resulted in a sample with 128,498 SNPs.

2.1 Performance Comparison of Current Methods

Our analysis considered each method's ability to correctly infer the degree of relatedness between the pairs of samples based on their reported relationships. These

reported relationships are extremely reliable and in most cases we can validate them via first degree connections among samples in the densely-genotyped SAMAFS pedigrees. Some methods directly infer the degree of relatedness[46] while others infer a kinship coefficient[1,48,50], a coefficient of relatedness[49,52] (which is two times the kinship coefficient[38]), or instead detect IBD segments[53–56] (Table 2.2). To infer the degree of relatedness from an estimated kinship coefficient for a pair of samples, we use the ranges of estimated kinship values from the KING method[1] (Table 1.1). These ranges use differences in powers of two for the relatedness degree intervals, which is generally consistent with simulations[40]. For IBD detection methods that report the number of IBD segments shared at a locus[53,56], it is straightforward to calculate a kinship coefficient[32]. This coefficient, ϕ_{ij}, between a pair of samples i, j denotes the probability that a randomly selected allele in individual i is IBD with a randomly selected allele from the same genomic position in j and is introduced in Section 1.2.3. For the IBD detection methods that do not distinguish between regions that are IBD1 from IBD2[54,55], the proportion of the genome that is inferred to be IBD0 provides an alternate means of estimating the degree of relatedness (Table 1.1), with the ranges of values here again from the KING paper[1]. We classified individuals with lower kinship coefficients or higher IBD0 rates than indicated for the fifth degree range as unrelated.

Using the SAMAFS sample, we assessed the performance of each program by using them to classify all pairs of individuals. Figure 2.2 shows the proportion of sample pairs inferred to be within each of the degree classes that we considered (first through

Method	Version	Citation Number	Type	Output	Parallelized?	Runtime (×cores used if >1)	Requires independent markers	Input required from outside program	Accounts for population structure
ERSA	2.0	46	IBD segment-based	Degree of relatedness	N	14.5h	N	IBD segments	NA
fastIBD	Beagle 3.3.2	54	IBD segment-finding	IBD segments	N	55.5h	N	NA	NA
GERMLINE (-haploid)	1.5.1	53	IBD segment-finding (Distinguishes IBD1 and IBD2)	IBD segments	N	20m	N	Phased genotypes	NA
IBDseq	r1206	55	IBD segment-finding	IBD segments	Y	33.5h (×16)	N	NA	NA
KING (KING-robust)	1.4	1	Allele frequency-based IBD estimate	IBD 0,1,2 proportions	N	5m	Y	NA	Y
PC-Relate	2.0.1	52	Allele frequency-based IBD estimate	IBD 0,1,2 proportions	N	9h	Y	Pairwise kinship coefficients	Y
PLINK 1.9	1.90b2k	49	Allele frequency-based IBD estimate	IBD 0,1,2 proportions	N	20s	Y	NA	N
PREST-plus	4.1	51	Allele frequency-based IBD estimate	IBD 0,1,2 proportions	N	179h	N	NA	N
REAP	1.2	48	Allele frequency-based IBD estimate	IBD 0,1,2 proportions	N	4h	Y	Ancestral population proportions	Y
Refined IBD	Beagle 4.1	56	IBD segment-finding (Distinguishes IBD1 and IBD2)	IBD segments	Y	91h (×16)	N	NA	NA
RelateAdmix	0.1	50	Allele frequency-based IBD estimate	IBD 0,1,2 proportions	Y	16h (×16)	Y	Ancestral population proportions	Y

Table 2.2: Properties of the 11 relationship inference methods we analyzed. Type indicates the inference methodology the program uses. Runtime is wall clock time to run the program; we ran parallelized programs using the numbers of cores indicated in parentheses: total compute time for the parallelized programs is the runtime multiplied by the number of cores used. Input required from outside program indicates extraneous information needed to run the program. Programs that use either principal components or ancestral population proportions are indicated as accounting for population structure. "Y" indicates yes, "N" indicates no, and "NA" indicates not applicable. Runtimes are from a machine with four AMD Opteron 6176 2.30 GHz processors (64 cores total) and 256 GB memory.

fifth degree and unrelated), with results separated according to the reported and inferred relatedness degrees of the pairs. All methods perform well when inferring first and second degree relatives, with the accuracy ranging from 98.4% to 99.5% for first degree relatives, and from 93% to 98.6% for second degree relatives. For more distant relatedness, the IBD-based methods have higher accuracy than those that rely on allele frequencies of independent markers—for example, for fifth degree relatives, the top performing IBD-based method has 59.4% accuracy while the highest performing allele frequency-based method has only 53.8% accuracy. Overall, the most accurate programs are ERSA 2.0, Refined IBD, and IBDseq. The improved accuracy of IBD-based methods may be due to their focus on identifying long stretches of identical segments that more readily discriminate recent shared relatedness from chance sharing of alleles.

Noting that the SAMAFS consist of admixed Mexican American individuals, we examined the accuracy results among the allele frequency-based methods, of which several account for population structure. Of all these methods, PC-Relate has the highest accuracy across all levels of relatedness, and it does account for population structure using principal components. Overall, the results are mixed with regards to accounting for population structure and accuracy, with PC-Relate, REAP, RelateAdmix, and KING all incorporating population structure into their models, and PREST-plus and PLINK ignoring this structure.

The inference accuracy of all methods decreases for higher relatedness degrees, likely

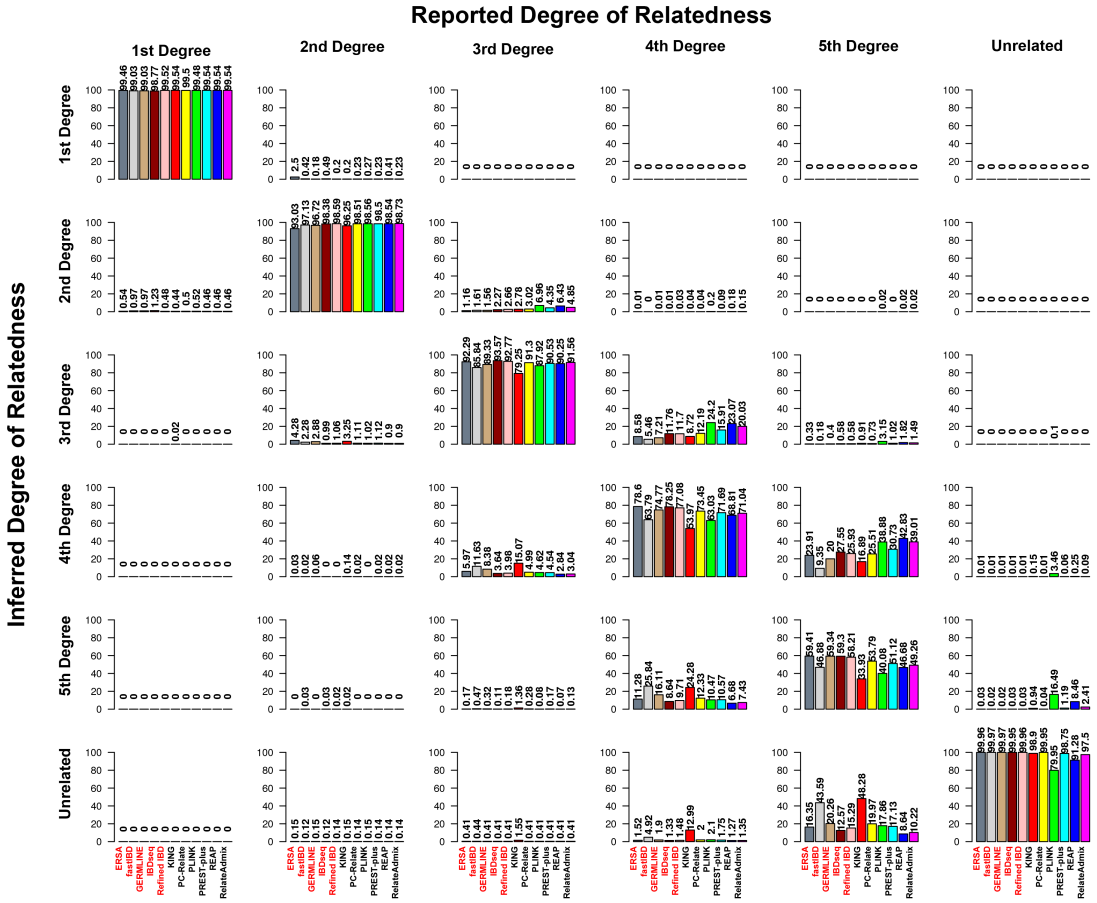

Figure 2.2: Performance comparison of the evaluated methods using the SAMAFS dataset. Bar plots indicate the percentage of pairs of samples that are reported to have a given degree of relatedness and who are inferred to be in each degree class. The bar plots are separated on the horizontal axis by the reported relatedness degree and on the vertical axis by inferred relatedness degree. For clarity, the plots list above each bar the percentage number that the corresponding bar depicts. Program names listed in red are IBD-based methods while those in black utilize allele frequencies for inference.

due to the exponential drop in mean pairwise IBD shared and an increased coefficient of variation as relatedness decreases[39,41,66]. In particular, for fifth degree relatives, the accuracy rates for all methods are very low at less than 60%. However, in nearly all cases ($\geq$ 83.8%), the programs correctly inferred the degree of relatedness to within one degree of the reported relationship. IBDseq has the highest within-one-degree accuracy for reported fourth degree pairs (the relationship class with the lowest accuracies for off-by-one inference) at 98.7%. At the same time, the methods classify an average of 97.9% of pairs of unrelated individuals correctly, averaged across all programs (99.7% when PLINK is excluded), with few instances of fifth or greater degree of relatedness inferred for these pairs. These results suggest that, when methods do detect relatedness—even as far distant as fifth degree—the individuals are likely to be truly related. As shown in Section 3.1, misreported or unknown relationships in the SAMAFS dataset likely explain some of the inference errors, particularly since even some confidently inferred first degree relationships were likely misreported as a more distant relationship. Overall, we find that IBD-based methods outperform other approaches for more distantly-related pairs, though notably these packages require substantially more compute time to run which may limit their utility in some applications (Table 2.2). While the precise performance results presented here are specific to the SAMAFS sample, we find that reducing the sample size to mimic a dataset that may not be as well-phased (Section 2.2.2) still produces similar results, with methods that leverage IBD segments having greater accuracy than other approaches. Therefore, the results presented here should be generalizable and indicate general properties of relationship inference methodologies: approaches

that use IBD segments outperform other methods for third degree and more distant relatives; and the specificity of relatedness inference, even in a dataset where phase accuracy may be relatively high, is inhibited for all but the closest relatives.

2.2 Accounting for Biases

As the SAMAFS data consist of numerous large families, allele and haplotype frequencies estimated from the sample may be biased, potentially affecting the inference results in a way that is not representative of the methods' accuracies in other datasets. Here, we attempt to address these questions and reanalyze relevant methods.

2.2.1 Allele Frequency Estimates

To assess whether potentially biased allele frequency estimates may impact the results of the allele frequency-based methods on the full data analysis (Figure 2.2), we tested PLINK[49] on datasets containing mostly unrelated individuals. To generate these sample sets, we first determined a set of unrelated individuals using FastIndep[67], a program that uses estimated kinship coefficients and a maximum allowed relatedness threshold to identify a set of individuals in which no pairwise relatedness exceeds the given threshold. For pairs reported as unrelated, we use the kinship

Degree	Number of Pairs
1	4,510
2	5,999
3	7,030
4	5,991
5	2,496
Unrelated	1,997,907
Total	2,023,933

Table 2.3: Numbers of pairs of individuals tested for each degree of relatedness for the analysis described in Section 2.2.1.

coefficients from PLINK, and for pairs reported as related, we use the expected kinship coefficient (Table 1.1) value for that pair. We input these kinship coefficients to FastIndep with the relatedness threshold set to 0.015 which is slightly below the lower bound for calling fifth degree relatedness (roughly 0.022). This produced a set of 529 individuals that have little to no genetic relatedness among them. We note that PLINK is somewhat biased in inferring relatedness and identifies a non-trivial proportion of samples that are reported to be unrelated as fifth degree or closer relatives (Figure 2.2). Therefore, using PLINK kinship estimates provides an aggressive filter against potential relatedness in these sample sets. Next, we created 1,000 datasets containing the base set of unrelated samples merged with no more than one randomly selected pair of related individuals from each SAMAFS pedigree, resulting in a total of 26,026 pairs of fifth degree or closer relatives and nearly two million unrelated pairs tested (Table 2.3). By adding only one randomly selected pair of related individuals from each pedigree, we limit the potential for bias. When adding

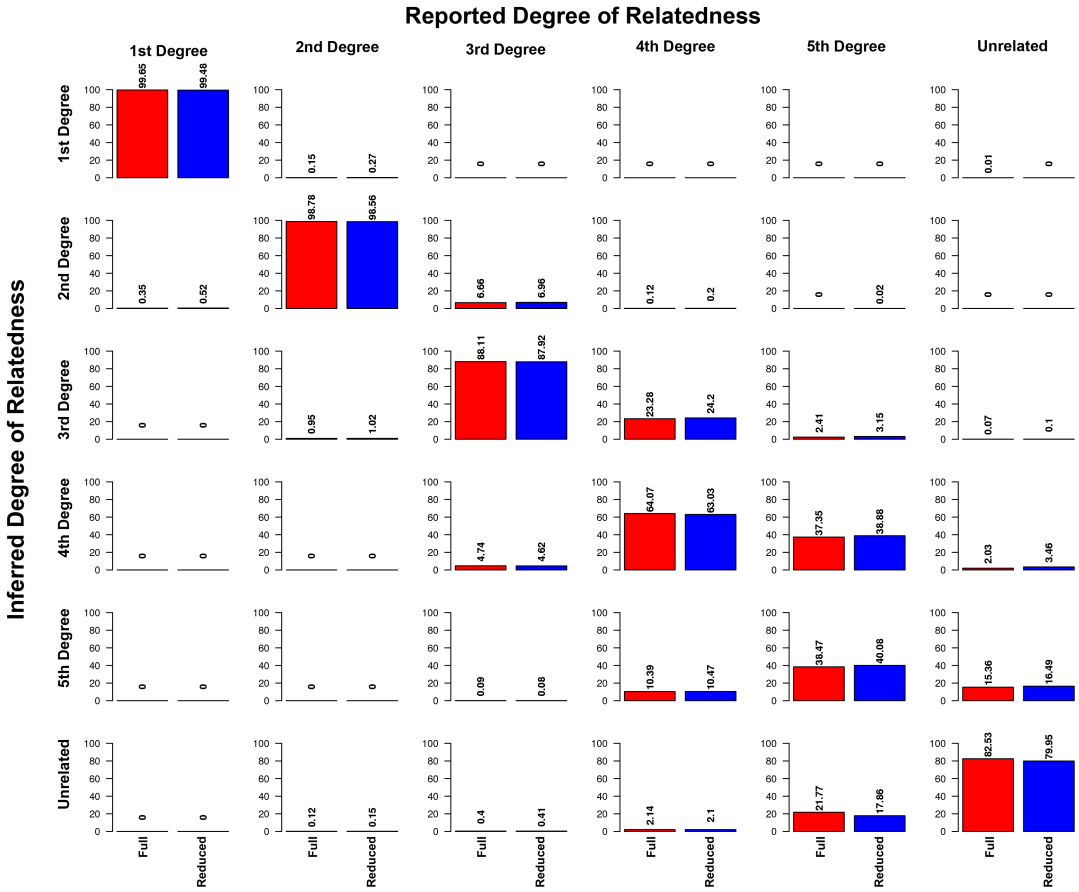

Figure 2.3: Accuracy results from PLINK run on the entire SAMAFS dataset denoted by red bars (labeled "Full") and from PLINK run on 1,000 reduced datasets composed of mostly unrelated individuals denoted by blue bars (labeled "Reduced").

a related pair of individuals to the dataset, we checked if either of the individuals was reported to be a fifth degree or closer relative of a sample in the set of unrelated individuals, and in that case, removed that previously unrelated individual from the dataset. Finally, we ran PLINK on each of the 1,000 datasets and show performance accuracy results in comparison to running PLINK on the full dataset in Figure 2.3.

While some differences exist between the two analyses, the accuracy results differ by less than 3% for all relatedness degrees, suggesting that allele frequency biases are small and only minimally impact inference accuracy.

2.2.2 Haplotype Phasing

Haplotype phasing and therefore IBD inference accuracy in this dataset of highly related individuals might be greater than would be achieved in a more outbred sample. This may increase the inference quality of IBD segment-detecting programs (which utilize either internal phasing models or pre-phased data) compared to the other programs. To assess the performance of the IBD segment-detection methods in a setting with relatively outbred data, we again used datasets comprised mostly of unrelated individuals. Specifically, starting with the 1,000 datasets generated as outlined above, we merged genotypes from 580 HapMap3 individuals (83 individuals of African ancestry in Southwest USA [ASW], 165 Utah residents with Northern and Western European ancestry from the CEPH collection [CEU], 77 samples of Mexican ancestry in Los Angeles, California [MXL], 88 Toscani in Italia individuals [TSI], and 167 Yoruba in Ibadan, Nigeria samples [YRI]) in order to increase the sample size. This provides a baseline level of phase accuracy that should be achievable for most studies as all these datasets contain between 1,127−1,204 individuals. Results from this analysis are presented in Figure 2.4. The accuracy of the IBD segment-based

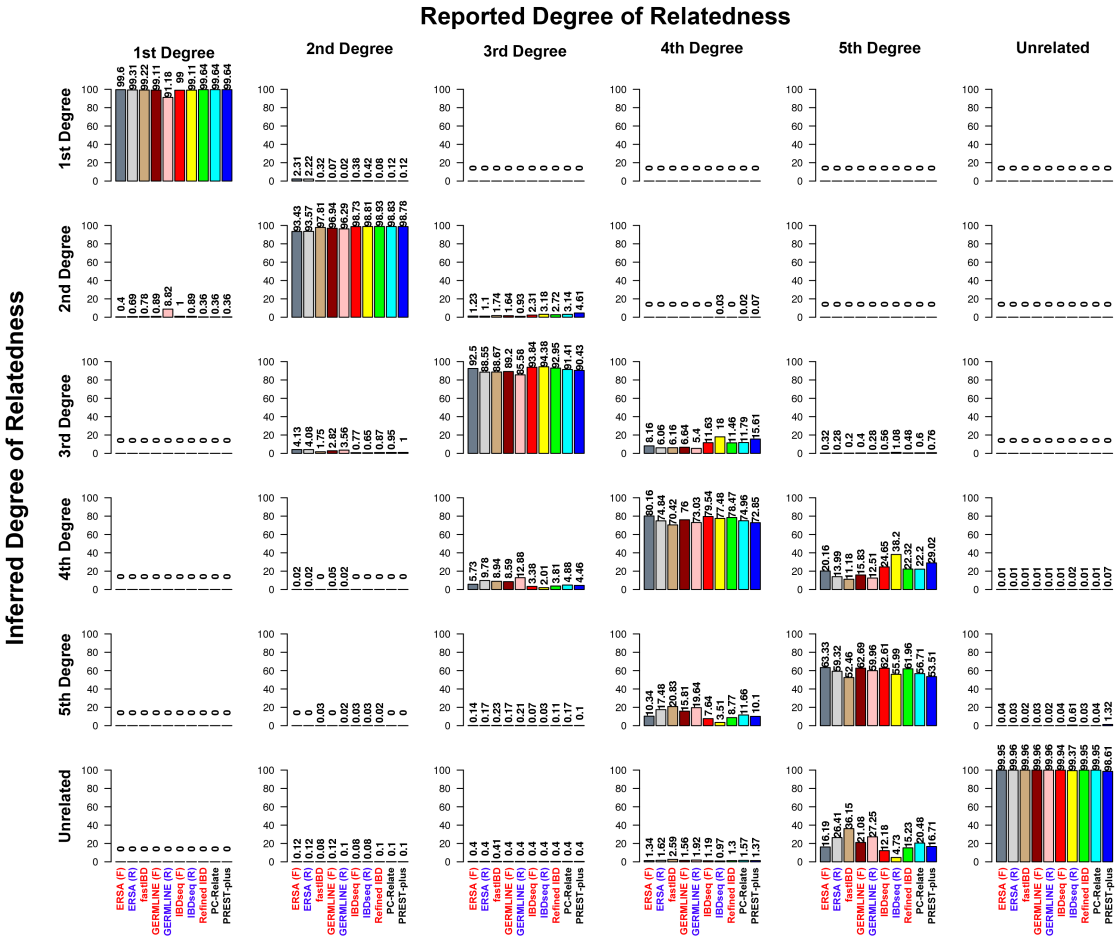

Figure 2.4: Accuracy results from the full dataset for all IBD-segment finding methods and PC-Relate and PREST-Plus along with results from running ERSA, GERMLINE, and IBDseq on the 1,000 reduced datasets. Results from programs run on both types of data are indicated with a label "(F)" and red text for the full dataset and "(R)" and blue text for the reduced datasets. The accuracies of all methods are for pairs of samples that were included in at least one reduced dataset so that the results are directly comparable between data types. When a pair of unrelated relatives is present in more than one reduced dataset, we randomly selected results from one program run on an arbitrary dataset to determine accuracy.

methods does drop for higher degrees of relatedness in the reduced datasets compared to all of SAMAFS, in some cases by as much as 8%. In this case the performance of IBD segment methods and allele frequency methods are more similar, suggesting that for smaller datasets, phasing errors can reduce the efficacy of IBD segment methods for inferring relatedness. Still, the IBD segment-based methods are comparable to or more accurate than the allele frequency methods even in this setting. Moreover, for larger datasets where it is possible to achieve phase accuracy at the megabase-scale[68], the results from the full dataset indicate that IBD segment finding methods provide greater accuracy than allele frequency methods for relatedness inference. This is true even in the reduced datasets that have no more than 1,204 samples and therefore are subject to a relatively high level of phasing errors.

2.2.3 Population Structure

Samples that have admixed ancestry can confound relatedness estimation methods due to the presence of admixture linkage disequilibrium, a genetic feature that induces an increased correlation in genotypes among admixed samples that are not recently related[48,69]. While methods such as REAP[48] and RelateAdmix[50] adjust for admixture, they rely on the output of model-based ancestry inference methods such as ADMIXTURE[70] which have difficulty distinguishing between ancestral populations and more recent relatedness among samples[69]. To ensure that the results

we obtained from ADMIXTURE (and used for REAP and RelateAdmix) represent population-level ancestry and not relatedness structure within the SAMAFS data, we ran ADMIXTURE in two ways. First, we introduced genotypic variance that corresponds to the desired population ancestry by generating a dataset containing the entire (LD-pruned) SAMAFS sample together with 372 unrelated HapMap3 individuals. These HapMap3 individuals are a subset of the 580 individuals described above (including samples with African, European, and Native American ancestry relevant to SAMAFS), but with samples filtered out by FastIndep using previously estimated kinship coefficients[71] as input and a filtering threshold of 0.015 as above. We then ran ADMIXTURE on this dataset with $K = 3$. Next, we ran ADMIXTURE with $K = 3$ on another dataset containing the 372 unrelated HapMap3 samples and the 529 SAMAFS samples believed to be unrelated (described above). This latter dataset has little relatedness structure and ADMIXTURE should therefore readily identify ancestral proportions for African, European, and Native American populations. Consistent with this, we located the ancestries likely to correspond to these groups by locating the ancestry coefficient with the highest values using individuals from the YRI, CEU, and MXL populations in the two different ADMIXTURE runs. We then computed correlations for the ancestral proportions inferred for each of these three components for the 529 unrelated SAMAFS samples contained in both ADMIXTURE runs. These correlations are extremely high at > 0.97 for all three populations, indicating that the output from ADMIXTURE run on all of SAMAFS together with the 372 unrelated HapMap3 individuals reliably infers population-level ancestry proportions. We therefore used these ADMIXTURE results (extracting only

the ancestry estimates for SAMAFS) for running REAP and RelateAdmix.

In contrast to REAP and RelateAdmix, which use input from a separate program to obtain ancestry proportions, PC-Relate[52] infers principal components itself on a set of unrelated individuals it locates in the data. As the authors note, a challenge arises in this context in determining how many principal components should be included to explain the population structure while not inadvertently discounting recent relatedness[52]. Still, PC-Relate performs well and was among the top methods that are based on allele frequencies.

CHAPTER 3

A COMPOSITE METHOD USING TOP-PERFORMING METHODS

As current methods provide only moderate accuracy when classifying third through fifth degree relatives, we evaluated the potential for increasing performance by combining inference results from the top three programs: ERSA 2.0, IBDseq, and Refined IBD. We used an approach that calls the degree of relatedness for a pair only when all three programs unanimously agree on the relatedness degree, providing no classification for other pairs. The resulting inference accuracy increased only negligibly (0.15%, 0.22%, 1.6%, 3.1%, 1.8%, and 0.01%, respectively for first through fifth degree and unrelated pairs) in comparison to the most accurate method's performance in each degree class. We also considered a majority vote between the three programs, discarding the cases in which all three programs inferred a different degree (only two cases were of this class). With this approach, there is a slight decrease in performance overall (-0.46%, -0.26%, -1.4%, -1.5%, +0.28%, +0.01%). These results suggest that while there is room for improvement in the specificity of relatedness inference methods, dramatic improvement is likely to be achieved only with novel approaches (Chapter 4) and not composites of current methods.

Here, we apply three pairwise methods in order to characterize relatedness in three datasets: SAMAFS, HapMap3, and a Qatari dataset. Although methods which aggregate results from various relatedness inference methods do not result in a great increase in accuracy, as a conservative measure we here apply this methodology. In

particular, we require unanimous agreement from our top three methods as determined in Section 2.1 (ERSA 2.0, IBDseq, and Refined IBD) for the relationships we infer.

3.1 Application to SAMAFS Data

We examined the pairs of samples that were inferred to be related but were reported as unrelated (in distinct pedigrees) in the SAMAFS dataset. ERSA 2.0, Refined IBD, and IBDseq all inferred a small number of first through third degree relationships that connect individuals from different pedigrees within SAMAFS (Figure 3.1). Numerous between-family relationships were discovered this way, with seven of the discovered relationships being first degree relationships (Table 3.1). Overall, we found 48 pairs of pedigrees with at least five pairs of first through third degree relatives between them which all three methods unanimously infer to have the same degree of relatedness. Additionally, these three methods agreed on the inference of 374 and 1,632 pairs of fourth and fifth degree relatives between the pedigrees (not shown). These results highlight the importance of checking for relatedness among samples in all cohorts, and indicate that there can be sizable numbers of relatives across a range of degrees even in well-studied samples.

We further searched for misreported relationships in the SAMAFS data, again look-

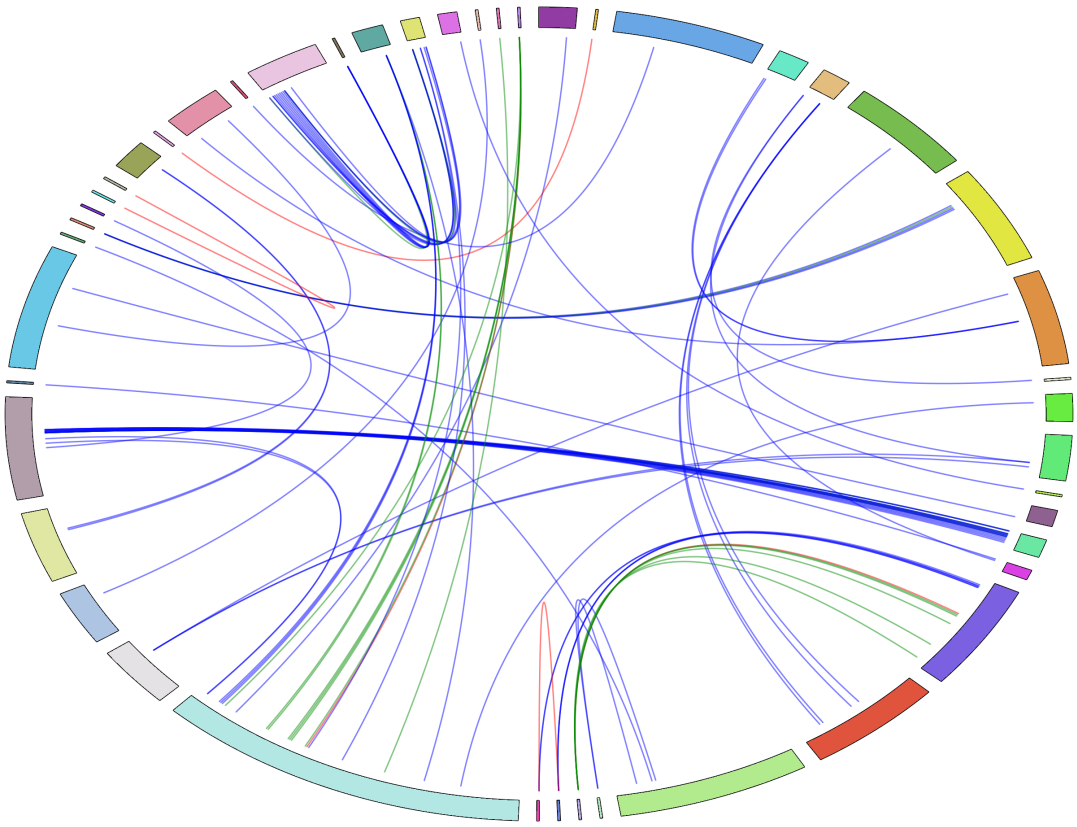

Figure 3.1: Relationships discovered between individuals from different SAMAFS pedigrees. Bands on the perimeter of the elliptical plot indicate distinct pedigrees within SAMAFS with band size proportional to the number of individuals in the pedigree. Curves between two bands correspond to discovered relative pairs with color indicating the degree of relatedness: red for first degree, green for second degree, and blue for third degree. Points where the curves end correspond to specific individuals, and a single point may have multiple curves running to it, indicating several relationships between that individual and others in the dataset.

ing at relationships in which all three methods unanimously agreed upon the degree of relatedness of the pair in question, but limited to relationships reported as first degree but inferred to be second degree or unrelated, and relationships reported as second degree but inferred to be first degree or unrelated. For inferred relationships that differ by more than one degree from the reported relationship (e.g., reported as second degree but inferred as unrelated or vice-versa), we assumed that the inference is valid as this is unlikely to occur due to data errors or statistical fluctuations. For relationships that are inferred to differ by only one degree from the reported relationship, we further required that either: (1) the discrepant relatedness call be supported by a consistent call involving at least one other sample (example follows); (2) in cases of reported siblings inferred to be second degree relatives, that their IBD2 proportion be less than $\frac{1}{2^{5/2}}$; or (3) in cases of reported half-sibling pairs inferred to be first degree relatives, that their IBD2 proportion be greater than $\frac{1}{2^{5/2}}$. As an example of an inference supported by another sample, given a set of three or more reported siblings, if the methods infer a pair of siblings as likely second degree relatives (presumably half-siblings), we checked that one of the other siblings also supports a second degree relatedness inference involving one of the two original samples to ensure consistency. We used Refined IBD's results to quantify IBD2 levels. Note that the expected proportion of IBD2 between full-siblings is $\frac{1}{4}$, and we used $\frac{1}{2^{5/2}}$ as the cutoff for confirming full- vs. half-siblings calls.

The IBD2 levels of two reported half-siblings from two pedigrees were greater than that seen for most half-siblings but less than typical for full-siblings, and appear to

		Inferred		
		1st	2nd	Unrelated
Reported	1st	**4,908**	23	0
	2nd (HS)	5	**388**	5
	2nd (A)	2	**4,789**	3
	2nd (GP)	0	**945**	0
	Unrelated	7	35	**3,023,456**

Table 3.1: Pairs of relationships that are confidently inferred using unanimous agreement from ERSA 2.0, IBDseq, and Refined IBD, and further checks described in the text (for some discrepant relationships) in SAMAFS. (HS) indicates half-sibling pairs, (A) indicates avuncular pairs, and (GP) indicates grandparent-grandchild pairs. Bolded numbers indicate the counts of agreements between the reported and inferred relationships. Pairs whose relationship were not unanimously agreed upon by the methods or which could not be verified as probable misreports using the checks we describe are not counted.

be best explained as being a less commonly described class of relatives known as three-quarter-siblings. Three-quarter-siblings are individuals who share one parent in common and whose unshared parents have a mean coefficient of relatedness of 50%—consistent with these parents being full-siblings. Individuals with this class of relatedness share non-trivial proportions of IBD2 but at a lower level than for full-siblings. For the potential three-quarter-siblings we identified, we did not have genotype data for one of the fathers in both cases and therefore could not validate whether the fathers were siblings. We note that we obtained reported relationships based on the SAMAFS pedigree structure which does not include information about the relationships between the two unshared parents of the reported half-siblings. Therefore, as these pedigrees indicate that the samples have only one parent in common, they are consistent with our observations and we did not consider them

discrepant and did not include them in Table 3.1.

3.2 Application to HapMap3 Data

We searched the HapMap3 dataset for unknown close relatives. Focusing only on pairs whose relationship class is unanimously agreed upon by the top three programs, we found several fifth degree relationships unreported in previous studies[29,71]: two in ASW (individuals with African ancestry in Southwest USA), six in LWK (Luhya in Webuye, Kenya), 67 in MKK (Maasai in Kinyawa, Kenya), and one in YRI (Yoruba in Ibadan, Nigeria) populations. The high level of discovered fifth degree relationships in the MKK population is consistent with the findings of Pemberton et al.[29] who suggest there may be considerable background relatedness in the sample, potentially due to certain cultural practices of marriage and reproduction[72,73] or due to recent demographic events affecting the Maasai population[74]. This high level of relatedness poses challenges to analyses of the demographic history of this population[75] and underscores the need to analyze relatedness in all genetic analyses.

3.3 Application to Qatari Data

We take the three top-performing programs according to our SAMAFS analysis and apply these to Qatari data collected by Weill Cornell Medicine in Qatar[76]. We make use of 108 genomes, each of which was sequenced to a median depth of 37 (minimum 30x) by Illumina technology.

Since estimated current rates of consanguinity in Qatar are around 22% of marriages with higher levels in the past[77], Qatari individuals are expected to show higher levels of inbreeding than those from other parts of the world. Previous analyses of the genomes of Qatari individuals found three distinct clusters reflecting differing ancestry[2,76,78]. One of these subpopulations, Qataris with Bedouin history, or those coined Q1 individuals in previous studies[2,76,78], presented larger inbreeding coefficients than the other subpopulations. We therefore expect to see a wider variance in estimated kinship coefficients (ϕ) for the Qatari individuals as consanguinity may result in higher levels of relatedness between affected pairs of individuals. This may introduce bias as the programs tested do not account for inbreeding. Although the kinship coefficient is intended to capture the effects of inbreeding on relatedness, it serves as a hindrance in cases like this where the level of inbreeding is unknown and one's main purpose is to classify relationships by a degree of relatedness rather than on a continuous scale. If parental relationships are known, one may input them into a modified version[79] of GENEHUNTER[80] which will output the expected pairwise

IBD0, IBD1, and IBD2 proportions between relatives, including in cases of inbreeding where at least one of the individuals being compared has parents that are related to one another. In our case, however, we have no prior information on whether parents may be related and, if so, what their relationship is.

We attempt to characterize the level of inbreeding in the individuals in our dataset by comparing the total runs of homozygosity (ROH) lengths of the Qataris using the exome dataset to the individuals in the SAMAFS dataset as shown in Figure 3.2. PLINK's `--homozyg` was used to calculate ROHs. On average, Qatari individuals have roughly nine times higher total ROH length than the SAMAFS individuals, suggesting that our inferred degrees of relatedness for pairs of Qatari individuals may be more error-prone due to the higher levels of inbreeding. However, the high level of background relatedness between the individuals in this dataset will likely improve phasing accuracy, suggesting that estimated kinship coefficients from IBD-based methods may not be as error-prone.

We find that in the Q1 subpopulation, we are able to classify most pairs of individuals as being second cousins or more closely related as expected due to the history of consanguinity in Qatar (Figure 3.3). Nodes denote Qatari individuals, colored by subpopulation, and lines between nodes indicate an inferred relationship of the indicated degree or more related. By the fourth degree, many lines form within the Q1 (red) subpopulation, indicating that there exist many closer relationships within this subpopulation and/or that the Q1 individuals indeed have higher levels of con-

sanguinity. None of the programs used account for consanguinity, and therefore we expect our inferred degrees of relatedness for the Qatari dataset to be biased. Even so, our results indicate that the methodologies are relatively sensitive to relatedness as they indicate a higher amount of genetic relatedness amongst individuals in the Q1 population than the other groups.

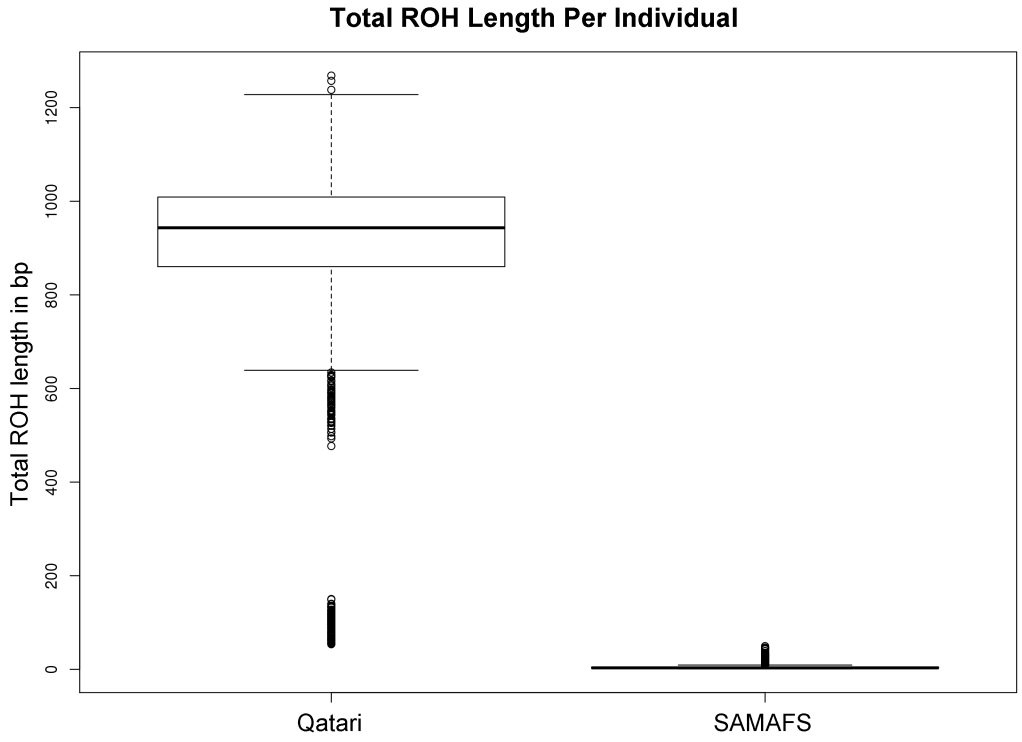

Figure 3.2: Total length (in base pairs) of runs of homozygosity in Qatari dataset versus SAMAFS dataset.

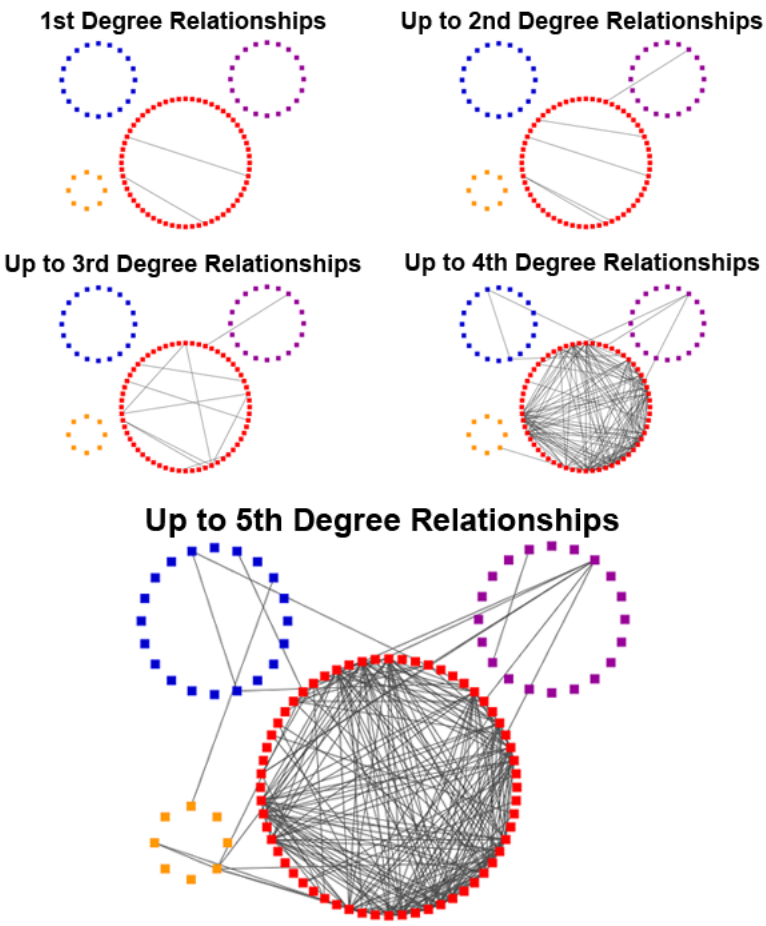

Figure 3.3: Relationships found between Qatari individuals up to given degree. Population labels Q1 through Q3 are described elsewhere[2]. Red nodes denote Q1 individuals, blue nodes denote Q2, purple nodes denote Q3, and orange nodes denote admixed. A line between two nodes indicate that a relationship was found between those two individuals at that degree of relatedness or more related.

CHAPTER 4

DRUID: DEEP RELATEDNESS UTILIZING IDENTITY BY DESCENT

Our earlier work (Chapter 2) and that of others show that relatedness inference accuracy declines as level of relatedness decreases: inference of first and second degree relatives is generally highly accurate, but performance decreases rapidly starting at third degree and continuing onward[30,46]. As the number of samples in a dataset increases, the number of expected relationships increases quadratically. With this increasing number of relationships, one can leverage information from sets of closely related samples to improve inference accuracy of relatedness between distant relative sets. We propose a method, Deep Relatedness Utilizing Identity by Descent, or DRUID, which effectively transforms the problem of inferring more distant relatedness to a problem of inferring a closer relationship: rather than infer relatedness between two individuals, it finds close relatives of the two individuals, and when possible, combines their IBD signals to infer the estimated relatedness between the ungenotyped parents or grandparents of these sets of close relatives. By using an individual's parent for inference rather than the individual, we reduce the problem of inferring a true degree of relatedness d to the problem of inferring a true degree of relatedness of $d-1$; similarly, by using an individual's grandparent, we reduce the problem to inferring a true degree of relatedness of $d-2$. By using both individuals' parents for inference rather than those individuals, we reduce the problem to a $d-1-1=d-2$ true degree of relatedness inference problem: each parent is

one degree closer to the other individual and hence two degrees closer to the other parent. Using both individuals' grandparents for inference reduces the problem to a $d-2-2=d-4$ true degree of relatedness inference problem as each grandparent is two degrees closer to the other individual and hence, the two grandparents are four degrees closer to one another than the original pair of individuals are to one another.

4.1 Method

DRUID uses input from an IBD detection algorithm to perform relatedness inference in two stages. First, it infers the pedigree structure of a set of close relatives who have a first degree relationship with at least one other sample—relationships that are very likely to be inferred correctly[30]. The method also infers and incorporates samples that are aunts and/or uncles of these first degree relatives using a new approach that leverages the fact that full-siblings share some genomic regions IBD on both haplotype copies. Second, DRUID combines IBD information from each set of close relatives to infer the expected genome-wide IBD sharing proportion between their ungenotyped ancestor and a more distant relative. Using this quantity, the method then infers the likely degree of relatedness between that ancestor and the distant relative, who may also be an ungenotyped individual. When the relationship to the distant sample arises through one of the close relatives' ancestors (not through

descent from one of these relatives), the ungenotyped ancestor will be more closely related to the distant sample than the genotyped individuals are. As relatedness inference accuracy is higher for closer relationships, i.e., for lower relatedness degrees[30], this approach provides greatly improved accuracy over pairwise relatedness methods that utilize data only from pairs of genotyped samples. We also show that it has improved accuracy over other methods that leverage the samples' relatedness structure to perform inference. Because most genotype datasets contain samples that were only collected relatively recently, we make the assumption that the distant relatives do arise through an ancestor of the close relatives and not via descent. We have implemented DRUID to utilize IBD segments detected using Refined IBD[56], but the approach is generally applicable to any method that reports whether samples share one or two IBD segments at a given position.

DRUID assumes there are no errors in the detected IBD segments and that there is no consanguinity. In particular, the two parents of a set of siblings are assumed not to be related to each other and all IBD segments are assumed to represent segments that are inherited from a common ancestor between two individuals. We have found that, although errors in detected IBD segments do occur, their overall effect is minor and combining all detected IBD segments extant in the descendants of an ungenotyped ancestor enables improved relatedness estimation.

4.1.1 Inferring Sets of Close Relatives

To infer the sets of closely related samples, DRUID generates a graph in which nodes represent samples and edge labels indicate the relationship type between the linked pair. The input IBD segments are informative about the relationships between the samples, and we use these to estimate the proportion of their genome that each pair of samples shares IBD either on one or two haplotype copies, denoted $\widehat{\text{IBD1}}$ and $\widehat{\text{IBD2}}$, respectively. These estimates are simply the sum of the lengths of the inferred IBD segments shared between the two samples on one or two haplotypes divided by the total genome length, with all lengths in cM units. From this, we derive the estimated kinship coefficients as $\hat{K} = \frac{1}{2} \times \widehat{\text{IBD2}} + \frac{1}{4} \times \widehat{\text{IBD1}}$ and deduce the likely relationship types based on the $\hat{K}$ and $\widehat{\text{IBD2}}$ values for each pair using the values in Table 4.1. Initially, the method considers only parent-child, full-sibling, and monozygotic (MZ) twin relationships.

Starting with an empty graph, DRUID adds nodes and edges corresponding to all inferred full-sibling relationships. Next, the method ensures that for all connected components, the nodes contained in it are all directly connected to one another as full-siblings. That is, if individual ind_1 was inferred to be full-siblings with ind_2 and ind_3, we ensure that ind_2 and ind_3 were also inferred to be full-siblings with one another. If one of the full siblings in a connected component does not have a full-sibling relationship with another individual in the component, DRUID removes

	$\widehat{\text{IBD2}}$	$\widehat{K}$
MZ Twin	$[\frac{1}{2^{1/2}}, 1]$	$[\frac{1}{2^{3/2}}, 1)$
Full Sibling	$[\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}})$	$[\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}})$
Parent-child	$[0, \frac{1}{2^{7/2}})$	$[\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}})$
2nd Degree	-	$[\frac{1}{2^{7/2}}, \frac{1}{2^{5/2}})$
3rd Degree	-	$[\frac{1}{2^{9/2}}, \frac{1}{2^{7/2}})$
4th Degree	-	$[\frac{1}{2^{11/2}}, \frac{1}{2^{9/2}})$
5th Degree	-	$[\frac{1}{2^{13/2}}, \frac{1}{2^{11/2}})$

Table 4.1: Relationship classification rules used by DRUID. The ranges of K and their mapping to relationships are those suggested by Manichaikul *et al.*[1] MZ twin: monozygotic twin.

the node with the fewest edges to the other nodes (in cases of ties, one of the nodes is selected at random), and continues this process until all pairs of nodes in each component are directly connected to one another. For any individuals that are pruned this way, we later add a generic first degree relationship edge between that individual and any sample previously inferred as its full-sibling. If MZ twins are present, we analyze only one of the samples and later report identical results for the omitted sample to those inferred for the analyzed twin.

Next, DRUID incorporates parent-child relationships into the graph, and, when possible, determines which individual is the parent using either (a) age information, (b) analysis of relatedness to full siblings, or (c) information provided by the user. Full sibling relationships provide information about which sample is the parent and which is the child in the following way. Suppose ind_1 and ind_2 are inferred to have

a parent-child relationship. If ind_1 has at least one full-sibling in the graph, then those full-siblings either all have a parent-child relationship to ind_2 (or general first degree relationship type), in which case ind_2 is the parent of all the full-siblings. Otherwise, ind_2 must be the child of ind_1 and have a second degree relationship to the full-siblings of ind_1. DRUID adds all inferred parent-child pairs to the graph, labeling which is the parent and which is the child when possible, and otherwise labeling them as a general first degree relative pair. When avuncular relationships are determined in this way, we add them to the graph as such, noting which individual in the pair is the aunt/uncle and which is the niece/nephew.

4.1.2 Incorporating Other Aunts and Uncles to the Set of Close Relatives

In principle, the length of IBD segments shared between second degree relatives are informative about their underlying relationship type: grandparent-grandchild, half-sibling, avuncular, or double cousins. However, the method RELPAIR, which implements an approach based on this idea, has limited ability to discriminate between these relationship types, with the classification accuracy ranging from 37% to 72% among these types[13] (excluding double cousins which the method does not consider). While analyses of IBD segment lengths between second degree relatives remains a promising direction, further work is needed to improve the inference resolution.

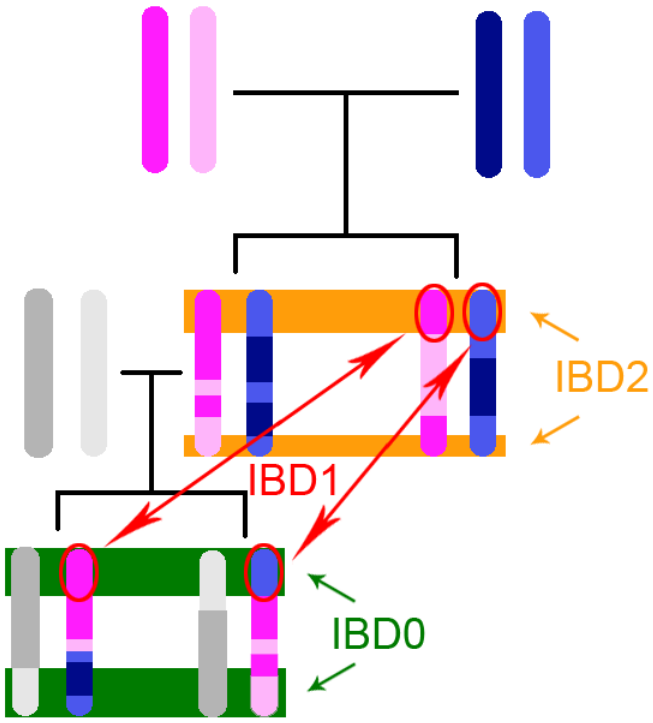

Figure 4.1: Haplotype transmissions in a pedigree with the relatedness structure indicated by black lines. The grandchildren (bottom haplotypes) are each IBD1 with their aunt/uncle at the top section of the chromosome (red ellipses) and are IBD0 with each other (green bars) in this region. Their parent is therefore IBD2 with the aunt/uncle (orange bars) at this locus. This scenario in which two siblings are IBD0 with each other and each are IBD1 to a given second degree relative suggests that the second degree relative is likely an aunt or uncle of the siblings.

In DRUID, we take a different approach based on the IBD sharing patterns among a set of three samples consisting of a pair of (full- or half-) siblings and a second degree relative, determining whether that relative is an aunt or uncle of the siblings. As described further below, siblings inherit distinct regions of each parent's genome, some regions identical to other siblings and some from different haplotype copies. The ungenotyped parent of a pair of siblings shares some regions IBD2 with the siblings' aunt or uncle, marking a unique sharing pattern that allows us to discriminate between second degree relatives and pinpoint aunts and uncles with high precision. In particular, at regions in which two siblings have inherited distinct haplotypes, which we detect as locations with no shared IBD segments between them or as IBD0, they will have inherited distinct haplotype copies from both parents. In these regions, if the two siblings both also share an IBD segment with a given second degree relative, one of their parents must share two distinct haplotypes IBD with that relative (ignoring double cousins in which both parents are related to the sample, a case we address below). Appreciable levels of this IBD pattern among the three samples are a strong indicator that the second degree relative is a full-sibling of the ungenotyped parent, or an aunt or uncle of the two siblings (Figure 4.1).

Following relative detection based only on first degree relative types (above), DRUID locates all inferred sets of close relatives containing two or more full-siblings for which data are unavailable for one or both parents. It then finds all samples that are inferred as a second degree relatives of each of these full-siblings and calculates the total genetic length of regions in which two of these siblings are IBD0 with each other

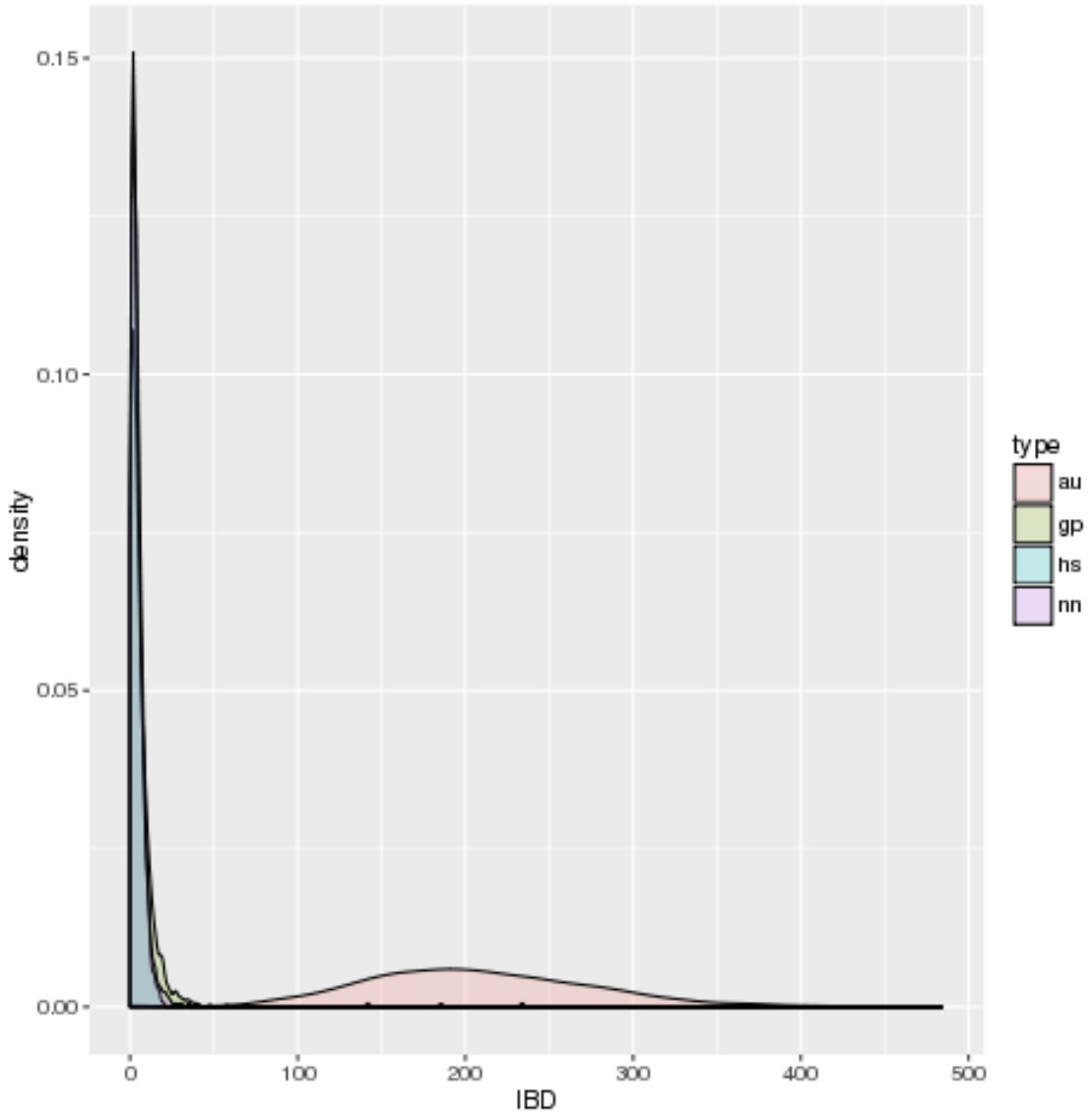

Figure 4.2: For each pair of siblings and an aunt/uncle, grandparent, or half-sibling of theirs in the set of trusted SAMAFS relationships (Section 4.2), we find regions in which the two siblings are IBD0 and are each IBD1 with the second degree relative, sum these regions, and plot the densities in the histogram. We do this for 2915 sets of a pair of siblings and their aunts/uncles, 970 sets of a pair of siblings and their grandparents, 731 sets of a pair of siblings and a half-sibling of theirs, and 595 sets of a pair of siblings and a niece/nephew of theirs. au: aunts/uncles; gp: grandparent; hs: half-sibling; nn: niece/nephew.

and both are IBD1 to the second degree relative. Our analyses indicate that when this pattern occurs in a total of >100 cM, the second degree relative is extremely likely to be an aunt or uncle of the siblings (Figure 4.2). If any two siblings have this level of the indicated sharing pattern with the second degree relative, he or she is added to the graph as an uncle or aunt with all siblings designated as a niece or nephew. We further include any siblings of this aunt or uncle as additional relatives of this same type, including the required relationship edges in the graph.

With the pedigree relationships between sets of close relatives inferred, DRUID can reconstruct the IBD profile of the ancestors of these sets. We focus on two types of close relative sets: full-siblings and full-siblings together with their aunts/uncles. We also show that using half-siblings provides accuracy results that are indistinguishable from inference using full siblings (Section 4.2.3). In order to make use of second degree relatives that are not an aunt/uncle of two or more siblings, we require a user to specify the relationships (including half-siblings). In the presence of such information, DRUID verifies that the samples are indeed second degree relatives and adds the relationship type edges to the graph.

4.1.3 Inferring IBD Sharing for a Parent Using Data from Siblings

A parent transmits to each child a random subset of the IBD segments he/she shares with any relative. Whereas a single child inherits only half of each parents' genome, data for additional children provide a more complete representation of the genomes of their parents, including receiving a larger fraction of the IBD segments they each carried. In particular, s-many full-siblings are expected to inherit a proportion of $Par(s) = 1 - \frac{1}{2^s}$ of both parents' genomes, a fact we exploit to infer the IBD sharing proportion of a parent given data for his/her children.

Given the assumption that the parents are unrelated, only one parent will have transmitted all IBD segments that a set of siblings share with any given distant relative D. Based on this assumption, although genetic data for the two parents are unobserved (Figure 4.3), the union of all IBD segments shared by the siblings with D constitutes a partial set of the IBD segments one of the parents shared with D. Notably, which parent transmitted these IBD segments is unknown, but this information is not needed to determine the degree of relatedness between that ungenotyped parent and D. Because we expect to observe a fraction $Par(s)$ of this parent's genome, and equivalently, that proportion of the genetic material the parent shared IBD with D, we compute the expected proportion of the genome this parent

P shared IBD on at least one haplotype copy with D as

$$\widehat{\text{IBD}}_{P,D} = \frac{\text{Length}\left(\bigcup_{c \in S} I_{c,D}\right)}{Par(|S|) \times L}. \tag{4.1}$$

Here, S is the set of siblings and $I_{c,D}$ is a set containing the markers that are called IBD between a given sibling (child) c and D. The Length(I) function gives the genetic length of all regions containing sequences of markers that are called IBD, and L is the total length of the genome, both in cM. As the union of all IBD regions in the siblings contains only an expected proportion $Par(|S|)$ of the parent's IBD regions, we scale the expected amount of genome shared IBD by the inverse of this quantity. This equation holds both for full-siblings and also for a combined set of full- and/or half-siblings: the expected proportion of the parent's genome present in such a set of individuals is also a function of its size (e.g., $s+h$ in Figure 4.4). The $\widehat{\text{IBD}}_{P,D}$ quantity maps directly to an estimated kinship coefficient as $\hat{K}_{P,D} = \frac{1}{4} \times \widehat{\text{IBD}}_{P,D}$, and we infer a degree of relatedness from this coefficient (Table 1.1).

An alternative to adjusting by the expected proportion $Par(s)$ of the parent's genome transmitted to the children is to estimate the actual transmitted proportion based on the observed IBD sharing between the siblings. Specifically, at positions where the children are all IBD2 with one another, both parents will have transmitted only one haplotype copy, or half of their genome. Likewise, at positions where at least two children are IBD0 with each other, each parent will have transmitted both haplotype copies or all their genetic material at these regions. We applied this logic to our analyses and compared the performance using these estimated proportions to using the expectation $Par(s)$. The results of both approaches are similar but using the

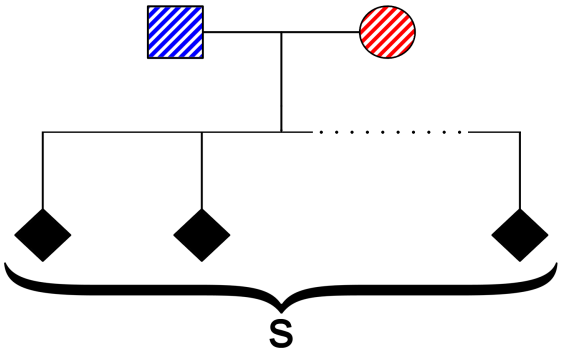

Figure 4.3: Reconstruction of the IBD profile between a distant relative and a parent more closely related to that relative than his/her children. Filled black individuals represent individuals for whom we have genotype data: here, s-many siblings. Individuals filled with stripes indicate the possible parents we can reconstruct the IBD profiles between themselves and the distant relative. We do not know which parent's IBD profile is being reconstructed.

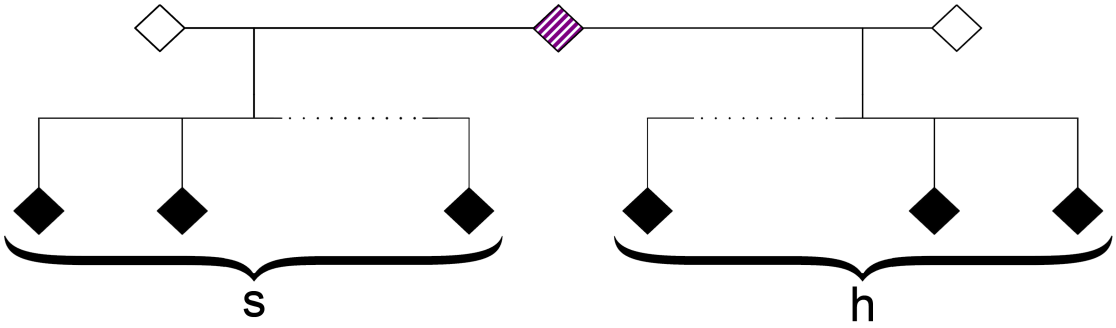

Figure 4.4: Reconstruction of the IBD profile between a distant relative and a parent more closely related to that relative than his/her children. Filled black individuals indicate individuals for whom we have genotype data: here, a set of s-many siblings and a set of h-many siblings, two sets of siblings that are half-siblings with one another. The individual filled with stripes is the parent whose IBD profile with the distant relative we reconstruct.

expectation yields somewhat higher accuracy, presumably owing to false negative or false positive IBD segments affecting the estimation (data not shown).

4.1.4 Inferring IBD Sharing for a Grandparent Using Siblings and Aunts/Uncles

When data are available for a set of siblings together with some number of their aunts and uncles, the IBD segments that these individuals share with a distant relative descend from a grandparent of the siblings and a parent of the aunts/uncles (Figure 4.5). The expected proportion of the grandparent's genome transmitted to these individuals is $Gr(k, s) = 1 - \frac{1}{2^k} + \frac{1}{2^{k+1}} \times Par(s)$, where k is the number of aunts/uncles of the s siblings. (This equation similarly holds when s is the number of full- and half-siblings included in the analysis.) Here, $1 - \frac{1}{2^k}$ is the expected amount of the grandparent's genome transmitted to the his/her k children and the final term gives the expected amount of DNA transmitted to $(k+1)^{\text{st}}$ child multiplied by the expected genome proportion that child—the parent of the siblings—transmitted to the s siblings.

With this expectation, we estimate the proportion of the genome that the grandpar-

ent G shares IBD with a distant relative D as

$$\widehat{\text{IBD}}_{G,D} = \frac{\text{Length}\left(\bigcup_{r\in(K\cup S)} \widehat{\text{IBD}}_{r,D}\right)}{Gr(|K|,|S|)\times L}, \tag{4.2}$$

where K is the set of aunts/uncles, S is the set of siblings, and L is again the genetic length of the genome under analysis.

As the sibling set may have aunts/uncles both through their mother and their father, we group together the aunts/uncles that are inferred to be siblings to create two sets of aunts/uncles, denoting these sets as K_{i_1} and K_{i_2} if so. If at least one set of aunts/uncles is available for each sibling set S_1 and S_2, we check whether

$$\sum_{k_1\in K_{1_i}, k_2\in K_{2_j}} \widehat{\text{IBD}}_{k_1,k_2} \times \frac{1}{|K_{1i}|+|K_{2j}|} > \sum_{s_1\in S_1, s_2\in S_2} \widehat{\text{IBD}}_{s_1,s_2} \times \frac{1}{|S_1|+|S_2|} \tag{4.3}$$

where $i \in \{1,2\}$ if S_1 has more than one aunt/uncle set, $i \in \{1\}$ otherwise, $j \in \{1,2\}$ if S_2 has more than one aunt/uncle set, $j \in \{1\}$ otherwise. If more than one pairing of $\{K_{1_i}, K_{2_j}\}$ fits this criteria, we use the pairing with the highest average estimated IBD. If no pairing fits this criteria, or if we have only one aunt/uncle set, we check whether the average of the estimated proportion of genome shared IBD with the relative(s) of interest and the aunts/uncles is at least as large as the maximum of estimated proportion of genome shared IBD with the relative(s) and each individual in the sibling set. In cases when there are two sets of aunts/uncles and both have higher average IBD shared with the relative than the maximum of that of the sibling set, we choose the aunt/uncle set with the higher average. When there is not a set of aunts/uncles that fit this criteria, we continue the analysis using only the sibling set.

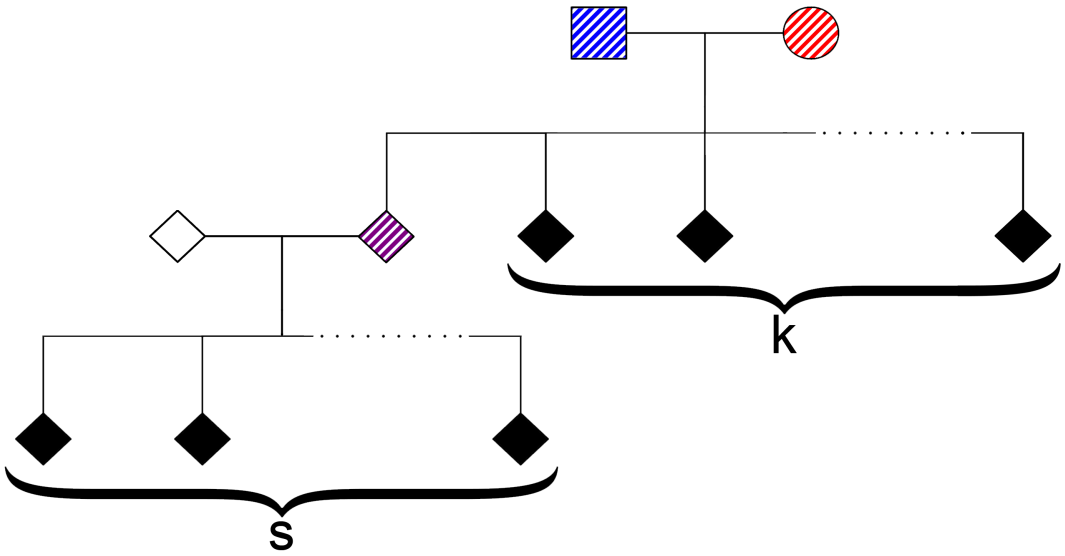

Figure 4.5: Reconstruction of the IBD profile between a distant relative and a grandparent more closely related to that relative than his/her grandchildren. Filled black individuals indicate individuals for whom we have genotype data: here, a set of s-many siblings and a set of their k-many aunts/uncles. The individual filled with purple stripes indicates the parent that is a sibling of the k-many aunts/uncles whose IBD profile with the distant relative we are able to reconstruct via the s-many siblings. The individuals filled with blue and red stripes indicate the possible grandparents whose IBD profiles with the distant relatives we reconstruct.

4.1.5 Estimation of More than One Parent's or Grandparent's IBD Profile

In sufficiently large datasets or those with family-based recruitment, DRUID will often infer several sets of closely related samples. In such cases, distant relatedness may exist between two sets of these close relatives and not merely to a single distant

relative d. In this case, inferring the amount of IBD shared between two ungenotyped ancestors from the two pedigrees enables inference at greater resolution than the potential alternative of using a single member of one of the pedigrees. Given two pedigrees with sets K_1, K_2 of aunts/uncles and S_1, S_2 of siblings, (with the sets of aunts/uncles allowed to be empty) we estimate the IBD sharing between the two ungenotyped ancestors a_1 and a_2 as

$$\widehat{\text{IBD}}_{a_1,a_2} = \frac{\text{Length}\left(\bigcup_{x\in R_1, y\in R_2} I_{x,y}\right)}{Anc(|K_1|,|S_1|) \times Anc(|K_2|,|S_2|) \times L}. \tag{4.4}$$

Here, $R_i = K_i \cup S_i$ for $i \in \{1,2\}$, and the expected proportion of the ancestor's genome transmitted to the corresponding set of close relatives is given by

$$Anc(k,s) = \begin{cases} Gr(k,s) & \text{if } k > 0 \\ Par(s) & \text{if } k = 0 \end{cases}.$$

4.1.6 Determining Relatedness across All Sample Pairs

To perform relatedness inference between all sample pairs, DRUID must determine which other members of any close relative sets to use for the inference and when to utilize standard pairwise relatedness measures. After its first stage of inferring the close relative sets, DRUID next infers a pairwise-only degree of relatedness for every two samples (above and Table 1.1). When this value is less than or equal to two for a given pair, DRUID reports that degree; additionally, if neither sample is in the

graph (i.e., neither is in a close relative set), DRUID reports the pairwise relatedness estimate. Otherwise, the method determines whether a parent or grandparent of a set of close relatives is in the graph, and if so, whether that ancestor has the same or higher genome proportion shared IBD with the other sample, successively moving up to older generations until arriving at two samples whose relatedness DRUID is to estimate.

Let i_1, i_2 be the two samples with relatedness to be inferred where at least one is a member of a set of close relatives. If neither i_1 nor i_2 have any siblings, half-siblings, or aunts/uncles, DRUID reports the pairwise degree of relatedness between these samples and deduces from this the relatedness degrees between them and all the descendants in the pedigrees to which they each belong. Let S_x denote the set containing x and his/her full-siblings (if any exist) for $x \in \{i_1, i_2\}$. If i_1 (likewise for i_2) has any half-siblings or aunts/uncles, DRUID checks whether they are likely related to i_2 through the same lineage as i_1, with half-siblings considered so if at least one half-sibling j has pairwise relatedness to $k \in S_{i_2}$ such that $\widehat{\text{IBD}}_{j,k} \geq \min_{a \in S_{i_1}, b \in S_{i_2}} \widehat{\text{IBD}}_{a,b}$. That is, we include half-siblings if at least one half-sibling has a pairwise IBD sharing proportion as large as the minimum relatedness between all pairwise IBD quantities between a full-sibling of i_1 and a full-sibling of i_2 (including i_1, i_2). For aunts/uncles of i_1 (likewise for i_2), we include them in the inference when they fit the criteria described in Section 4.1.5 or Section 4.1.4. Based on the identified collection of informative siblings and aunts/uncles, DRUID performs inference using one of Equations 4.1–4.4).

4.2 Accuracy of DRUID

To assess our method, we used SNP array genotypes from Mexican American individuals contained in large pedigrees from the San Antonio Mexican American Family Studies (SAMAFS)[57–59]. We describe this dataset in the Section 2.1, "Performance Comparison of Current Methods". Our analysis in Sections 2.1 and 3.1 show that there may be some misreports and/or unreported relationships in the SAMAFS dataset, and we therefore rely on the results of our method based on Refined IBD[30]: we merge the results of three runs of Refined IBD using different random seed values and for each pair of individuals, determine which regions of the genome they share IBD1 or IBD2, calculate the proportion of the genome shared IBD1 or IBD2 by dividing by the genetic length of the genome, and from this, calculate estimated kinship coefficients and inferred degrees of relatedness. When a reported sibling pair was not estimated to have a degree of relatedness of one by our Refined IBD method, we discarded that sibling pair. In our analysis of DRUID's performance, we compare its results to that of the Refined IBD method.

To ensure we trust the reported aunts/uncles of these verified sibling sets, for each reported aunt/uncle, we test whether all siblings are inferred to be second degree relatives of that aunt/uncle by Refined IBD. If so, we accept this individual as an aunt/uncle. As second degree inference has a slightly lower accuracy than first degree inference for Refined IBD, for each aunt/uncle verified in this manner, we

check whether he/she has any verified siblings but which were not inferred to be second degree relatives of initial sibling set: if so, we add these individuals as an aunt/uncle of the initial sibling set.

For reported half-siblings of a set of full-siblings, we ensure that the inferred relationship between the each of the reported half-siblings and each individual in the initial set of siblings is second degree according to the Refined IBD method. Similar to the aunt/uncle verification process, if a verified set of siblings are reported to be half-siblings to another verified set of siblings, we require all siblings in one set be inferred as second degree relatives of at least one sibling in the other set. When this occurs, we keep all verified siblings in each sibling set and label the pairwise relationships between the two sibling sets as half-sibling, otherwise we keep both sibling sets but do not label any pair as half-sibling.

4.2.1 Accuracy Using Sibling Sets

To enable direct comparisons of inferences using different numbers of full-siblings, we restrict our analysis to sibling sets with five or more individuals, yielding a total of 45 sets of siblings. We ignore any reported non-full-sibling relationships in this analysis.

For each set of siblings, we find all relatives such that all siblings are reported by pedigree to be third, fourth, or fifth degree relatives to those individuals. Thus, we ensure that the relative is not a descendant of a single sibling. We then perform inference on each included set of full-siblings (where $s \geq 5$) and these distant relatives by first randomly sampling two of the siblings and inferring their relatedness using DRUID. We next randomly sample another sibling which was not yet sampled, add him/her to the set of two siblings, and again infer relatedness between them and all relevant relatives. We continue this until we have sampled and tested five siblings together. Thus, for each sibling set, the siblings included in smaller numbers of siblings are subsets of those for larger numbers of siblings. We do this to make the inferences for the different sibling set sizes more directly comparable. If the full-sibling set is of size ten or larger, we repeated this process on the siblings that were not yet sampled.

We find an overall trend of increasing accuracy as the number of siblings s increases for third, fourth, and fifth degree relationships as shown in Figure 4.6. As degree of relatedness increases, so does DRUID's gain in accuracy in comparison to the Refined IBD method, with the largest gain in accuracy, 15%, found in the case of fifth degree inference when $s = 5$. Even when we have only a single pair of siblings, we see a considerable increase in accuracy: 2.4% for third degree, 7.7% for fourth degree, and 9.5% for fifth degree.

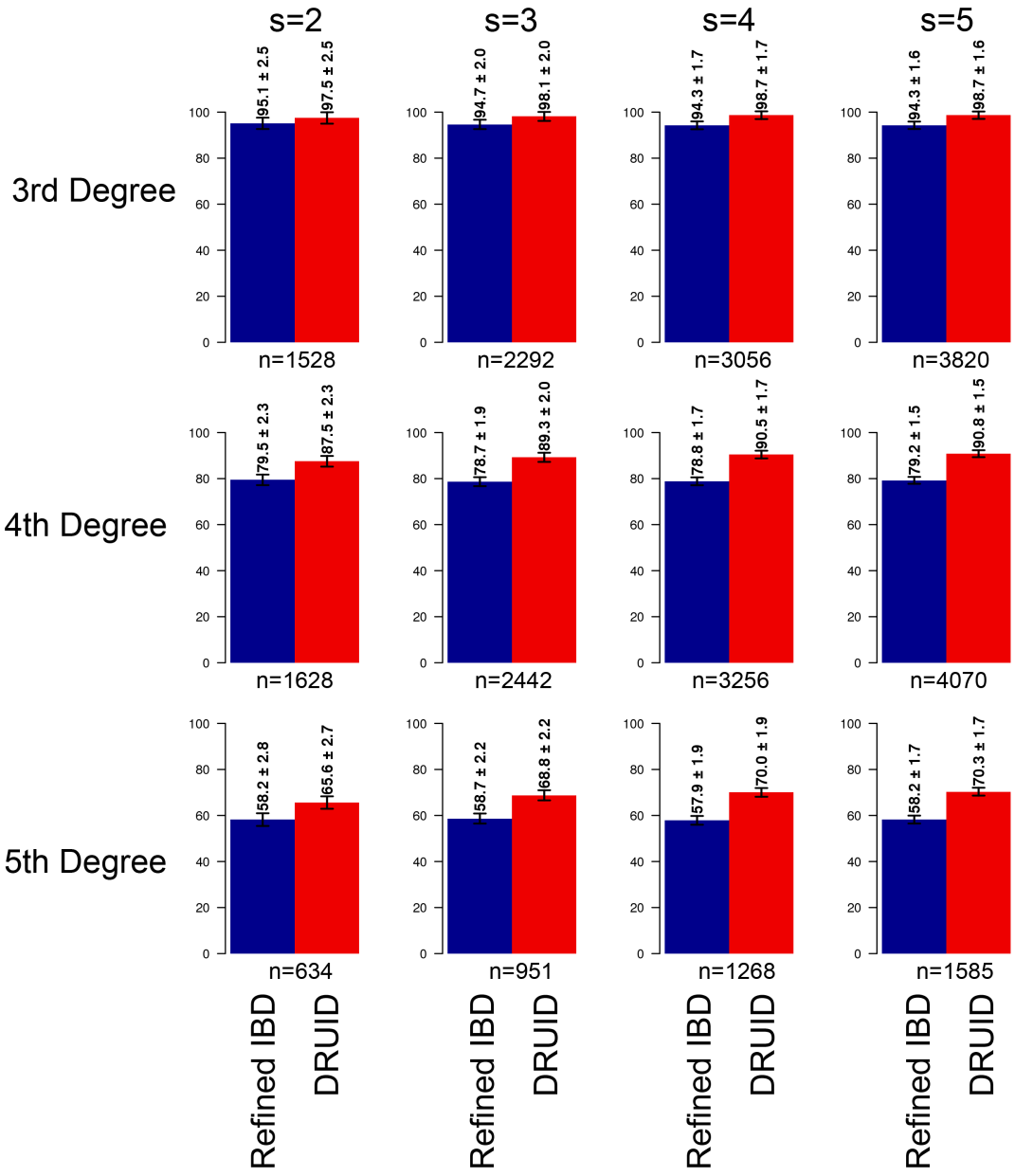

Figure 4.6: Results from the sibling-only analysis. s indicates the number of siblings included. n indicates the total number of pairs of individuals for which we obtain results: in the case of $s = 2$, $n = 1528$ for third degree, meaning 764 sets of a pair of siblings and a third degree relative were compared. Blue bars indicate the Refined IBD-based method's results, red bars indicates DRUID's results. Error bars denote 95% confidence intervals which were generated by bootstrapping 1000 samples.

4.2.2 Accuracy Using Siblings and Their Aunts/Uncles

For each set of siblings, we find all verified aunts/uncles. We find all relatives to which all aunts/uncles are reported by pedigree to be equally related and who are third, fourth, or fifth degree relatives of these aunts/uncles. We further check that each sibling in the sibling set is reported to be one degree further in relatedness than the aunt/uncle set to these relatives. We only consider individuals who are reported to be third and fourth degree relatives of the aunt/uncle set, as parties reported to be third degree relatives of the initial sibling set will be second degree relatives of the aunt/uncle set, and DRUID immediately reports such relationships. Given a set of siblings and aunts/uncles of sizes $s \geq 5$ and $k \geq 2$, respectively, we randomly sample two siblings and infer their relatedness between only these two siblings all relevant relatives, again using the same process as in the sibling-only analysis to test two, then three, four, and five siblings at a time. We then randomly select one aunt/uncle and repeat the same testing scheme in the sibling-only analysis but including this aunt/uncle. Finally, we randomly select a second aunt/uncle and repeat the same testing scheme in the sibling-only analysis but including both sampled aunts/uncles. We further include the case of $s = 1$ by randomly selecting a sibling from the various $s = 2$ cases and carrying out inference between this sibling and one or two of his/her aunts/uncles. Thus, for each sibling set and their aunts/uncles, the siblings included in smaller numbers of siblings are subsets of those for larger numbers of siblings and the aunts/uncles included in the smaller numbers of aunts/uncles are subsets of

those for larger numbers of aunts/uncles. Again, we do this to make the inferences for the different sibling set sizes and different aunt/uncles set sizes more directly comparable. If we have a sibling set of size 10 or larger with an aunt/uncle set of size four or larger, we repeat this entire process, sampling from individuals that have not yet been sampled.

Figure 4.7 shows our results with accuracy calculated using only inferences between the individuals in youngest generation (the s siblings) and the distant relatives, suggesting the considerable effect on accuracy when aunts and uncles related to the distant relative through the same lineage are included. When even one avuncular pair is known, we see a 11.6% increase in accuracy for fourth degree and a 15.8% increase in accuracy for fifth degree. When avuncular pairs are included, we are essentially estimating the s sibling's grandparent's IBD profile with the given relative via the k aunts/uncles and the ungenotyped parent of the s siblings. For example, in the case of $s = 2$ and $k = 1$ for fifth degree relationships, we see an accuracy of 74.4% which is within the 95% confidence interval surrounding the 80.6% accuracy of fourth degree relatedness for $s = 2$ and $k = 0$.

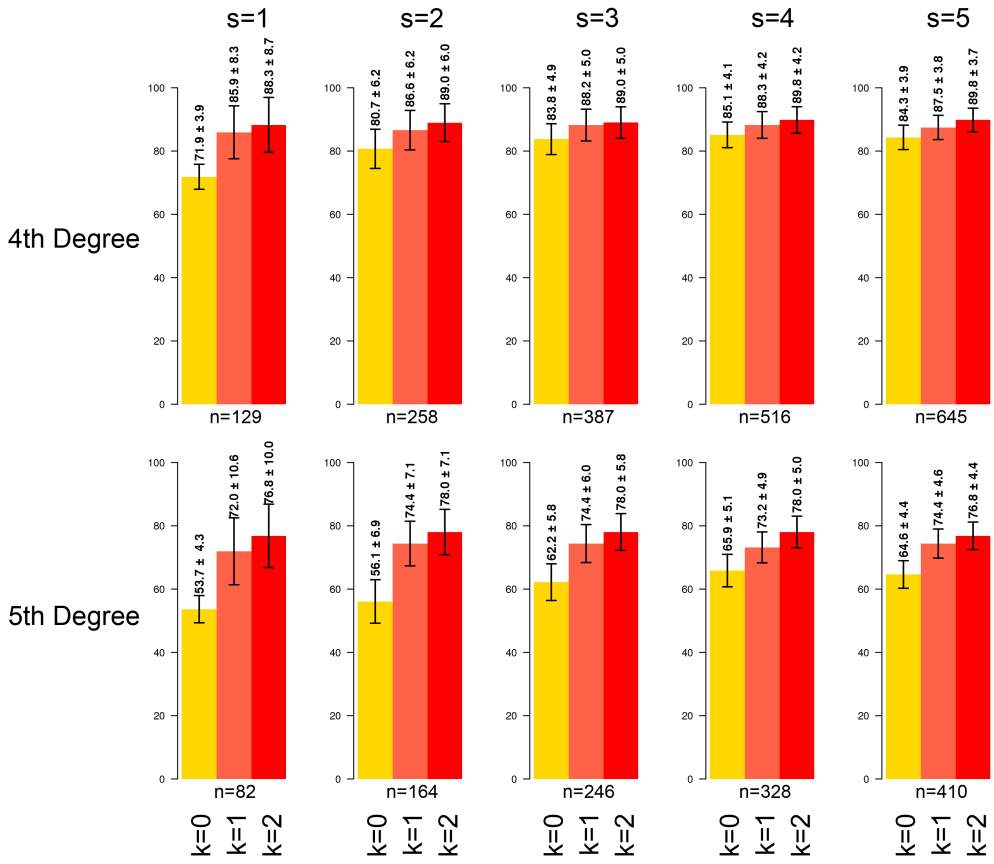

Figure 4.7: Results from the avuncular analysis. Degrees of relatedness are between the sibling set in the youngest generation and the distant relative. s indicates the number of siblings included, k indicates the number of aunts/uncles of those siblings included. n indicates the total number of pairs of individuals for which we obtain results that involve an individual from the base generation (the sibling set): in the case of $s = 2$, $n = 258$ for fourth degree, meaning 159 sets of a pair of siblings and a fourth degree relative were compared. As it is not possible to combine any IBD information in the $s = 1$, $k = 0$ case, we report the accuracy of the Refined IBD method as this is what DRUID falls back on in such case. Error bars denote 95% confidence intervals which were generated by bootstrapping 1000 samples.

4.2.3 Accuracy Using Half-Sibling Sets

DRUID uses both full-siblings and half-siblings in its inference. In principle, half-siblings provide the same amount of information about distant relatives of their parents as an equal number of full-siblings do. To test DRUID's ability to leverage half-siblings, for a sibling set of size five or larger, we find relatives to which all siblings in a sibling set are reported to be equally related and that are reported to be third, fourth, or fifth degree relatives of the siblings. If there are any verified half-sibling sets of this initial sibling set and these half-sibling sets are of at least size two, we determine to which distant relatives they are each also reported to be equally related as the initial sibling set. For each of these relatives, we find all relevant half-sibling sets. We take the largest set of siblings (between the initial sibling set and the half-sibling set(s)) and let that sibling set be what we refer to as the main sibling set. We randomly sample three siblings from the main sibling set and two half-siblings (full-siblings of one another) from the half-sibling set. We initially carry out analysis between all three siblings and distant relatives relevant to the half-sibling set using the Refined IBD method and DRUID. We randomly replace one of the three siblings with a randomly selected half-sibling from the two sampled half-siblings, and carry out analysis between the two full-siblings and their half-sibling. We randomly replace a second of the three siblings using the remaining sampled half-sibling and carry out analysis between the remaining individual in the main sibling set and two of his/her half-siblings. We then start again with the original three siblings, but remove the

sibling that was never randomly selected for replacement, carrying out analysis with the remaining two siblings. The first half-sibling previously randomly selected to replace a sibling then once again replaces the full-sibling he/she previously replaced, and analysis is carried out with one individual from the main sibling set and one individual from the half-sibling set. If we have a main sibling set of size five or larger and either a half-sibling set of size four or larger or two half-sibling sets of size two or larger, we randomly select two siblings that were not previously sampled from the main sibling set and another two half-siblings not previously sampled from the half-sibling set if possible, switching to the next half-sibling set and sampling from there if not possible.

Our results as shown in Figure 4.8 suggest that the inclusion of half-siblings is effective: when a more distant relative is related to a set of siblings and half-siblings through the same lineage, the accuracy using half-siblings appears to be the same as using full-siblings (with statistical fluctuations due to randomization and low sample size). Our results are also consistent with those from the sibling-only analysis: when half-siblings who are also related to the relative are included in the parent's reconstruction, we have the same trend of the accuracy increasing as the total number of siblings (and half-siblings) increases.

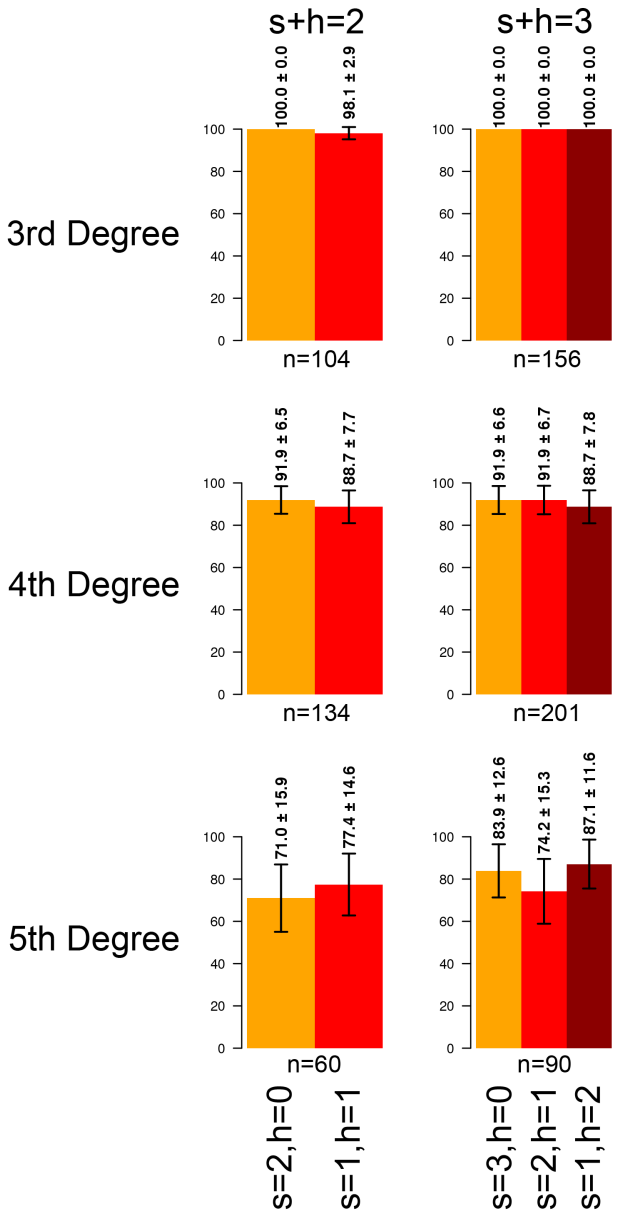

Figure 4.8: Results from the half-sibling analysis. s indicates the number of siblings included, h indicates the number of half-siblings included. n indicates the total number of pairs of individuals for which we obtain results: in the $s = 2$ case, $n = 166$ for third degree, meaning 83 sets of a pair of siblings (or half-siblings for the n=1 and h=1 case) and a third degree relative were compared. Error bars denote 95% confidence intervals which were generated by bootstrapping 1000 samples.

4.3 Comparison to PADRE

PADRE[81] is a method for inferring relatedness between two inferred pedigree structures. Specifically, PADRE makes use of pedigrees reconstructed by PRIMUS[82] which is a pedigree reconstruction method that takes genome-wide IBD proportions inferred using PLINK and computes likelihoods of relationship types to generate possible pedigrees of first, second, and third degree relatives. PADRE combines the output from PRIMUS with that of ERSA[60], an IBD segment-based method that uses IBD segments inferred by other programs such as GERMLINE[53] and, similar to PRIMUS, uses likelihoods to infer relatedness, but reportedly accurately infers relatedness up to 9th degree[46]. PADRE attempts to use ERSA-generated relationship likelihoods to identify the highest composite likelihood connection between family networks reconstructed by PRIMUS.

We find all sibling sets in SAMAFS that were verified (as described in Section 4.2). For each family with at least one sibling set, if more than two sibling sets are available, we analyze each pair of sibling sets that is reported to be third, fourth, or fifth degree relatives of one another, and infer relatedness between all individuals in the two sibling sets. If not, or after having done this, we then for each sibling set find all third, fourth, and fifth degree relatives that do not have siblings available, and infer relatedness between the sibling set and their relative. To test PADRE, we feed PRIMUS the results of PLINK as described in Section 2.1, but limited to

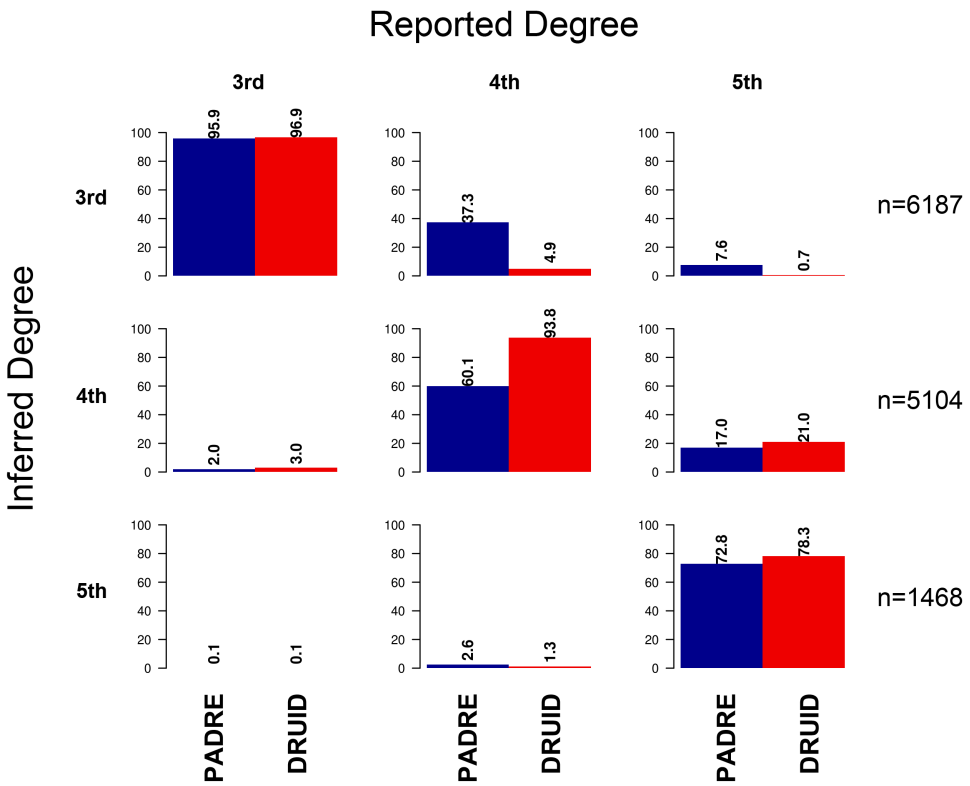

Figure 4.9: Comparison of PADRE (blue) and DRUID (red) using sets of verified siblings (Section 4.2) and their reported third, fourth, and fifth degree relatives. When a relative of a sibling set has siblings available, we use the method described in Section 4.1.5 to reconstruct the IBD profile of two ancestors; otherwise, we use the method described in Section 4.1.3 to reconstruct the IBD profile of only one ancestor. Barplots at the (inferred degree x, reported degree x) positions of the plot represent the true positive rates of the methods.

the individuals in the pair of sibling sets or in the set of siblings and their single relative, and run it with the default parameters. We input these PRIMUS results and the ERSA 2.0 results as described in Section 2.1 to PADRE and use the default parameters.

DRUID outperforms PADRE at third, fourth, and fifth degree inferences by 1.4%, 33.7%, and 5.5%, respectively (Figure 4.9). We find that although ERSA has a high accuracy rate[30], PRIMUS's third degree relative inferences tend to be biased upward, possibly due to inflated PLINK kinship coefficient estimates (Section 2.1). Since PADRE's results are highly dependent on that of PRIMUS, PADRE similarly is biased upward.

CHAPTER 5

SUMMARY AND CONCLUDING REMARKS

Relatedness inference is a key component of several forms of analyses, such as association studies, linkage analysis, and population genetics, where incorrectly accounting for relatedness or ignoring relatedness between samples can result in spurious and biased signals[11–13,18–20]. It also plays a role in forensic studies where it can allow for the identification of victims of disaster, relatives of missing persons, or criminals[14–16]. Further, relatedness inference has caught the attention of the general public as it is a fundamental tool to aid in the discovery of one's ancestry and genealogy. Companies such as 23andMe and AncestryDNA advertise their ability to connect their customers to others to whom they are likely related, marking a new era in the effort to reconstruct individuals' genealogical relationships.

Since before 1922[38], geneticists have pushed to understand and characterize relatedness between individuals. Identity by descent, or IBD, is a means to finding and understanding that relatedness. By estimating the percent of the genome that individuals likely inherited from the same common ancestor, it is possible to infer their degree of relatedness. When estimates of the proportion of genome shared IBD2 are available, one may infer the kinship coefficient of two individuals, as well as differentiate between first-degree relative types or determine whether any recent consanguinity occurred[34,35]. IBD sharing can be estimated via allele frequency-based similarity measures or haplotype-based similarity measures, with numerous methods

of both types available.

The age of big data is enabling the field of genetics to perform genetic studies with unprecedented accuracy. Though we are now capable of making amazing discoveries previously unreachable thanks to the collection of massive datasets, the sizes of existing datasets and ongoing studies requires more meticulous scrutiny of relationships of the samples to each other. In this thesis, we have shown that misreported or unreported close relatives within genotypic datasets can occur even in long-term studies that included extensive quality control measures. These errors can cause biased and spurious signals in various types of analyses, and therefore we stress the necessity of both verifying reported relationships and checking for unreported relationships.

We tested 11 state-of-the-art methods for inferring relatedness between individuals (Table 2.2). These methods were either allele frequency-based or IBD segment-based, and they output estimated kinship coefficients[1,48–50,52], degree of relatedness[46], or IBD segments[53–56]. We applied these methods to 2,485 Mexican American individuals in the SAMAFS dataset genotyped at 521,184 SNPs within pedigrees that span up to six generations with genotype data from as many as five generations of individuals. Given this large sample, including 13 pedigrees with >50 individuals (Figure 2.1), we extracted thousands of first through fifth degree relationships as well as millions of unrelated relationships (Table 2.1) via an in-house script and analyze all these pairs using the 11 methods.

We find that overall, all methods perform well when inferring first and second degree relatives, with the accuracy ranging from 98.4% to 99.5% for first degree relatives, and from 93% to 98.6% for second degree relatives (Figure 2.2). However, for more distant relatedness, their accuracy falls precipitously when classifying third to fifth degree relatives. This is unsurprising given the increased coefficient of variation as well as greater skewness in the proportion of genome shared as the meiotic distance between two relatives increases. Despite these challenges, the inferred relationship was within one degree of the reported relationship at a rate of $83\% - 99\%$ for all programs and relationship degrees (Figure 2.2). IBD segment-based methods—particularly, ERSA 2.0, IBDseq, and Refined IBD—outperform allele frequency-based methods, even when accounting for the increased phasing accuracy in the SAMAFS dataset due to the large number of closely-related individuals (Figure 2.4). We believe that the improved accuracy of IBD-based methods may be due to their focus on identifying long stretches of identical segments that more readily discriminate recent shared relatedness from chance sharing of alleles. We further find that all methods classify an average of 97.9% of pairs of unrelated individuals correctly, averaged across all programs (99.7% when PLINK is excluded), with few instances of fifth or greater degree of relatedness inferred for these pairs. These results suggest that, when methods do detect relatedness—even as far distant as fifth degree—the individuals are likely to be truly related.

We applied the three top-performing methods from our relatedness analysis (ERSA 2.0, Refined IBD, and IBDseq) to three datasets—SAMAFS, HapMap3, and Qatari

data collected by Weill Cornell in Qatar—in attempt to find unreported relationships. In SAMAFS, we checked for unreported relationships that all three methods unanimously agreed on, finding eight first degree, 20 second degree, 402 third degree, 374 fourth degree, and 1,632 fifth degree pairs that were unreported. Further, we find cases of likely unreported three-quarter-siblings, or individuals who share one parent in common and whose unshared parents have a mean coefficient of relatedness of 50%—consistent with these parents being full-siblings. In our analysis of the HapMap3 individuals, we find the three methods unanimously agree on several previously unreported fifth degree relationships, especially in the MKK population which is consistent with previous findings[29] and suggests that there many be considerable background relatedness in the sample due to certain cultural practices of marriage and reproduction. Similar to the MKK population, our analysis of the Qatari data reveals a high number of unreported relationships in the Q1 subpopulation of Qatar, consistent with previous findings of high levels of consanguinity in that subpopulation[2,76,78]. We therefore believe one should be careful in the analysis of datasets consisting of populations with high levels of consanguinity such as the Qatari dataset as the background relatedness between members of the population is likely higher than that of non-consanguineous populations. Overall, our discovery of unreported relationships ranging from first degree to fifth degree in all three datasets emphasizes the need to check for unreported relationships in all datasets.

Finally, we have developed the novel method DRUID which combines signals from multiple closely related samples to improve inference accuracy of relatedness between

distant relative sets. For two distantly-related individuals, i and j, for whom we wish to infer a degree of relatedness, DRUID first finds sets of relatives closely related to each of those individuals (relationships we can infer with high accuracy according to our results in Section 2.1) and combines information across all these close relatives to reconstruct the IBD profile between ancestors of i and j — ancestors who are more closely related to one another than i and j. This essentially reduces our problem of inferring a true degree of relatedness d to a problem of inferring a degree of relatedness $d - k$ for some $k > 0$, greatly improving accuracy. Together with this relatedness inference method, we devised a new approach for inferring aunts/uncles of a set of two or more siblings. This method leverages the fact that there are non-trivial amounts of IBD2 between the ungenotyped parent of the sibling set and the aunt/uncle of the sibling set which we are able to infer based on IBD sharing between siblings and the aunt/uncle. When using this approach to infer aunts and uncles of a set of siblings, we apply DRUID to reconstruct the IBD profile of the grandparents of the siblings (parents of the aunts/uncles). We find that DRUID outperforms both Refined IBD and PADRE, two state-of-the-art methods for inferring relatedness: when just two siblings are available for analysis, DRUID's accuracy for third, fourth, and fifth degree relationship inferences surpasses that of Refined IBD by 2.4%, 8%, and 7.4%, respectively; when five siblings and two of their aunts/uncles are available, DRUID's accuracy for fourth and fifth degree relationship inferences (degree with respect to the sibling set) surpasses that of Refined IBD by 17.9% and 23.1%, respectively. In the comparison to PADRE, we find that DRUID outperforms PADRE at third, fourth, and fifth degree relationship inferences by 1.4%, 33.7%, and 5.5%, respectively.

As datasets grow, the proportion of samples that have at least one relative in a dataset is expected to grow quadratically. With the increasing number of close relatives, DRUID's potential to improve inference accuracy will grow as well, as it leverages these numerous samples to more fully characterize the complete relatedness structure of the individuals under study. Thus, DRUID is poised to become an essential method in the current era of big data and personalized medicine. When sample sizes eventually reach millions of individuals, DRUID and extensions of it will allow the inference of hundreds of small to moderately sized pedigrees. This potential offers a glimpse toward a time when relatedness inference may encompass all samples in one very large pedigree structure that captures the historical relationships of all individuals to each other.

BIBLIOGRAPHY

[1] Ani Manichaikul, Josyf C Mychaleckyj, Stephen S Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, 2010.

[2] Larsson Omberg, Jacqueline Salit, Neil Hackett, Jennifer Fuller, Rebecca Matthew, Lotfi Chouchane, Juan L Rodriguez-Flores, Carlos Bustamante, Ronald G Crystal, and Jason G Mezey. Inferring genome-wide patterns of admixture in Qataris using fifty-five ancestral populations. *BMC Genetics*, 13(1): 1, 2012.

[3] U.S. Department of Health and U.S. Department of Energy Human Services. Understanding our genetic inheritance. the U.S. Human Genome Project: The first five years. *Government Printing Office, Washington, DC*, 1990.

[4] Richard A Gibbs, John W Belmont, Paul Hardenbol, Thomas D Willis, Fuli Yu, Huanming Yang, Lan-Yang Ch'ang, Wei Huang, Bin Liu, Yan Shen, et al. The international HapMap project. *Nature*, 426(6968):789–796, 2003.

[5] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[6] Nicholas J Schork, Sarah S Murray, Kelly A Frazer, and Eric J Topol. Common vs. rare allele hypotheses for complex diseases. *Current opinion in genetics & development*, 19(3):212–219, 2009.

[7] Vivien Marx. Biology: The big challenges of big data. *Nature*, 498(7453): 255–260, 2013.

[8] Jonathan Marchini, Lon R Cardon, Michael S Phillips, and Peter Donnelly. The effects of human population structure on large genetic association studies. *Nature Genetics*, 36(5):512–517, 2004.

[9] Joel N Hirschhorn and Mark J Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108, 2005.

[10] Benjamin F Voight and Jonathan K Pritchard. Confounding from cryptic relatedness in case-control association studies. *PLOS Genetics*, 1(3):e32, 2005.

[11] Jeffrey R O'Connell and Daniel E Weeks. PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *American Journal of Human Genetics*, 63(1):259–266, 1998.

[12] Jurg Ott. *Analysis of human genetic linkage*. JHU Press, 1999.

[13] Michael P Epstein, William L Duren, and Michael Boehnke. Improved inference of relationship for pairs of individuals. *American Journal of Human Genetics*, 67(5):1219–1231, 2000.

[14] Mark A Jobling and Peter Gill. Encoded evidence: DNA in forensic analysis. *Nature Reviews Genetics*, 5(10):739–751, 2004.

[15] Bruce S Weir, Amy D Anderson, and Amanda B Hepler. Genetic relatedness analysis: modern data and new challenges. *Nature Reviews Genetics*, 7(10): 771–780, 2006.

[16] Manfred Kayser and Peter de Knijff. Improving human forensics through advances in genetics, genomics and molecular biology. *Nature Reviews Genetics*, 12(3):179–192, 2011.

[17] Frederick R Bieber, Charles H Brenner, and David Lazer. Finding criminals through DNA of their relatives. *SCIENCE-NEW YORK THEN WASHINGTON-*, 5778:1315, 2006.

[18] David C Queller and Keith F Goodnight. Estimating relatedness using genetic markers. *Evolution*, pages 258–275, 1989.

[19] Laurence D Hurst. Genetics and the understanding of selection. *Nature Reviews Genetics*, 10(2):83–93, 2009.

[20] Joshua G Schraiber and Joshua M Akey. Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics*, 16(12):727–740, 2015.

[21] Dina L Newman, Mark Abney, Mary Sara McPeek, Carole Ober, and Nancy J Cox. The importance of genealogy in determining genetic associations with complex traits. *The American Journal of Human Genetics*, 69(5):1146–1148, 2001.

[22] Lon R Cardon and Lyle J Palmer. Population stratification and spurious allelic association. *The Lancet*, 361(9357):598–604, 2003.

[23] William Astle and David J Balding. Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, pages 451–471, 2009.

[24] Yoonha Choi, Ellen M Wijsman, and Bruce S Weir. Case-control association testing in the presence of unknown relationships. *Genetic epidemiology*, 33(8): 668–678, 2009.

[25] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8): 904–909, 2006.

[26] Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463, 2010.

[27] Margaret A Hamburg and Francis S Collins. The path to personalized medicine. *New England Journal of Medicine*, 363(4):301–304, 2010.

[28] Peter Ralph and Graham Coop. The geography of recent genetic ancestry across Europe. *PLoS Biol*, 11(5):e1001555, 2013.

[29] Trevor J Pemberton, Chaolong Wang, Jun Z Li, and Noah A Rosenberg. Inference of unexpected genetic relatedness among individuals in HapMap Phase iii. *American Journal of Human Genetics*, 87(4):457–464, 2010.

[30] Monica Ramstetter, Thomas D Dyer, Donna M Lehman, Joanne E Curran, Ravindranath Duggirala, John Blangero, Jason G Mezey, and Amy L Williams. A performance assessment of relatedness inference methods using genome-wide data from thousands of relatives. *bioRxiv*, page 106013, 2017.

[31] G Mendel. Versuche u ber planzen-hybriden. verhandlungen des naturforschenden vereines in brunn, bd. iv for das jahr 1865, abhandlungen, 3–47. *Genetic Theory*, 295:3–47, 1866.

[32] Elizabeth A Thompson. Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics*, 194(2):301–326, 2013.

[33] Bruce S Weir, Lon R Cardon, Amy D Anderson, Dahlia M Nielsen, and William G Hill. Measures of human population structure show heterogeneity among genomic regions. *Genome Research*, 15(11):1468–1476, 2005.

[34] A Jacquard. The genetic structure of populations. Translated by CHARLESWORTH and CHARLESWORTH from Structure geniques des populations, 1974.

[35] EA Thompson. The estimation of pairwise relationships. *Annals of Human Genetics*, 39(2):173–188, 1975.

[36] Brook G Milligan. Maximum-likelihood estimation of relatedness. *Genetics*, 163 (3):1153–1167, 2003.

[37] Amy D Anderson and Bruce S Weir. A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. *Genetics*, 176(1): 421–440, 2007.

[38] Sewall Wright. Coefficients of inbreeding and relationship. *The American Naturalist*, 56(645):330–338, 1922.

[39] WG Hill and BS Weir. Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genetics Research*, 93(01):47–64, 2011.

[40] Doug Speed and David J Balding. Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics*, 16(1):33–44, 2015.

[41] Peter M Visscher. Whole genome approaches to quantitative genetics. *Genetica*, 136(2):351–358, 2009.

[42] Mary Sara McPeek and Lei Sun. Statistical tests for detection of misspecified relationships by use of genome-screen data. *American Journal of Human Genetics*, 66(3):1076–1094, 2000.

[43] Lei Sun, Kenneth Wilder, and Mary Sara McPeek. Enhanced pedigree error detection. *Human Heredity*, 54(2):99–110, 2002.

[44] Sofia Kyriazopoulou-Panagiotopoulou, Dorna Kashef Haghighi, Sarah J Aerni, Andreas Sundquist, Sivan Bercovici, and Serafim Batzoglou. Reconstruction of genealogical relationships with applications to Phase iii of HapMap. *Bioinformatics*, 27(13):i333–i341, 2011.

[45] Eric L Stevens, Greg Heckenberg, ED Roberson, Joseph D Baugher, Thomas J Downey, and Jonathan Pevsner. Inference of relationships in population data using identity-by-descent and identity-by-state. *PLOS Genetics*, 7(9):e1002287, 2011.

[46] Hong Li, Gustavo Glusman, Hao Hu, et al. Relationship estimation from whole-genome sequence data. *PLOS Genetics*, 10(1), 2014.

[47] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81 (3):559–575, 2007.

[48] Timothy Thornton, Hua Tang, Thomas J Hoffmann, Heather M Ochs-Balcom, Bette J Caan, and Neil Risch. Estimating kinship in admixed populations. *American Journal of Human Genetics*, 91(1):122–138, 2012.

[49] Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1):1, 2015.

[50] Ida Moltke and Anders Albrechtsen. RelateAdmix: a software tool for estimating relatedness between admixed individuals. *Bioinformatics*, 30(7):1027–1028, 2014.

[51] Lei Sun and Apostolos Dimitromanolakis. PREST-plus identifies pedigree errors and cryptic relatedness in the GAW18 sample using genome-wide SNP data. *BMC Proceedings*, 8(Suppl 1):S23, 2014.

[52] Matthew P Conomos, Alexander P Reiner, Bruce S Weir, and Timothy A Thornton. Model-free estimation of recent genetic relatedness. *American Journal of Human Genetics*, 98(1):127–148, 2016.

[53] Alexander Gusev, Jennifer K Lowe, Markus Stoffel, Mark J Daly, David Alt-

shuler, Jan L Breslow, Jeffrey M Friedman, and Itsik Pe'er. Whole population, genome-wide mapping of hidden relatedness. *Genome Research*, 19(2):318–326, 2009.

[54] Brian L Browning and Sharon R Browning. A fast, powerful method for detecting identity by descent. *American Journal of Human Genetics*, 88(2):173–182, 2011.

[55] Brian L Browning and Sharon R Browning. Detecting identity by descent and estimating genotype error rates in sequence data. *American Journal of Human Genetics*, 93(5):840–851, 2013.

[56] Brian L Browning and Sharon R Browning. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, 194(2): 459–471, 2013.

[57] Braxton D Mitchell, Candace M Kammerer, John Blangero, Michael C Mahaney, David L Rainwater, Bennett Dyke, James E Hixson, Richard D Henkel, R Mark Sharp, Anthony G Comuzzie, et al. Genetic and environmental contributions to cardiovascular risk factors in Mexican Americans. *Circulation*, 94(9): 2159–2170, 1996.

[58] Ravindranath Duggirala, John Blangero, Laura Almasy, Thomas D Dyer, Kenneth L Williams, Robin J Leach, Peter O'Connell, and Michael P Stern. Linkage of type 2 diabetes mellitus and of age at onset to a genetic location on chromosome 10q in Mexican Americans. *American Journal of Human Genetics*, 64(4): 1127–1140, 1999.

[59] Kelly J Hunt, Donna M Lehman, Rector Arya, Sharon Fowler, Robin J Leach, Harald HH Göring, Laura Almasy, John Blangero, Tom D Dyer, Ravindranath Duggirala, et al. Genome-wide linkage analyses of type 2 diabetes in Mexican Americans. *Diabetes*, 54(9):2655–2662, 2005.

[60] Chad D Huff, David J Witherspoon, Tatum S Simonson, Jinchuan Xing, W Scott Watkins, Yuhua Zhang, Therese M Tuohy, Deborah W Neklason, Randall W Burt, Stephen L Guthery, et al. Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome Research*, 21(5):768–774, 2011.

[61] Amy L Williams, Giulio Genovese, Thomas Dyer, Nicolas Altemose, Katherine Truax, Goo Jun, Nick Patterson, Simon R Myers, Joanne E Curran, Ravi Duggirala, et al. Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *eLife*, 4:e04637, 2015.

[62] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[63] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[64] Giulio Genovese, Robert E Handsaker, Heng Li, Eimear E Kenny, and Steven A McCarroll. Mapping the human reference genomes missing sequence by three-way admixture in Latino genomes. *The American Journal of Human Genetics*, 93(3):411–421, 2013.

[65] International HapMap 3 Consortium et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, 2010.

[66] William G Hill. Variation in genetic identity within kinships. *Heredity*, 71: 652–653, 1993.

[67] Kuruvilla Joseph Abraham and Clara Diaz. Identifying large sets of unrelated individuals and unrelated markers. *Source code for biology and medicine*, 9(1): 1, 2014.

[68] Po-Ru Loh, Pier Francesco Palamara, and Alkes L Price. Fast and accurate long-range phasing in a UK Biobank cohort. *Nature Genetics*, 2016.

[69] Matthew P Conomos, Michael B Miller, and Timothy A Thornton. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genetic Epidemiology*, 39(4):276–293, 2015.

[70] David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664, 2009.

[71] Eric L Stevens, Joseph D Baugher, Matthew D Shirley, Laurence P Frelin, and Jonathan Pevsner. Unexpected relationships and inbreeding in HapMap phase iii populations. *PLOS ONE*, 7(11):e49575, 2012.

[72] Ernestina Coast. Maasai marriage: a comparative study of Kenya and Tanzania. *Journal of comparative family studies*, pages 399–419, 2006.

[73] Aud Talle. serious games: licences and prohibitions in Maasai sexual life. *Africa*, 77(03):351–370, 2007.

[74] Thomas Spear, Richard Waller, et al. *Being Maasai: ethnicity and identity in East Africa*. James Currey Publisher, 1993.

[75] Pier Francesco Palamara, Todd Lencz, Ariel Darvasi, and Itsik Peer. Length distributions of identity by descent reveal fine-scale demographic history. *American Journal of Human Genetics*, 91(5):809–822, 2012.

[76] Juan L Rodriguez-Flores, Khalid Fakhro, Francisco Agosto-Perez, Monica D Ramstetter, Leonardo Arbiza, Thomas L Vincent, Amal Robay, Joel A Malek, Karsten Suhre, Lotfi Chouchane, et al. Indigenous Arabs are descendants of the earliest split from ancient populations. *Genome Research*, 2016.

[77] AL Sandridge, J Takeddin, E Al-Kaabi, and Y Frances. Consanguinity in Qatar: knowledge, attitude and practice in a population born between 1946 and 1991. *Journal of Biosocial Science*, 42(01):59–82, 2010.

[78] Haley Hunter-Zinck, Shaila Musharoff, Jacqueline Salit, Khalid A Al-Ali, Lotfi Chouchane, Abeer Gohar, Rebecca Matthews, Marcus W Butler, Jennifer Fuller, Neil R Hackett, et al. Population genetic structure of the people of Qatar. *American Journal of Human Genetics*, 87(1):17–25, 2010.

[79] Anne-Louise Leutenegger, Emmanuelle Génin, Elizabeth A Thompson, and Françoise Clerget-Darpoux. Impact of parental relationships in maximum lod score affected sib-pair method. *Genetic Epidemiology*, 23(4):413–425, 2002.

[80] Leonid Kruglyak, Mark J Daly, Mary Pat Reeve-Daly, and Eric S Lander. Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics*, 58(6):1347, 1996.

[81] Jeffrey Staples, David J Witherspoon, Lynn B Jorde, Deborah A Nickerson, Jennifer E Below, Chad D Huff, University of Washington Center for Mendelian Genomics, et al. PADRE: Pedigree-aware distant-relationship estimation. *The American Journal of Human Genetics*, 99(1):154–162, 2016.

[82] Jeffrey Staples, Dandi Qiao, Michael H Cho, Edwin K Silverman, Deborah A Nickerson, Jennifer E Below, University of Washington Center for Mendelian Genomics, et al. PRIMUS: rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. *American Journal of Human Genetics*, 95(5): 553–564, 2014.